Figure 25.5 shows that the independent predictions are centered around the true values (i.e., errors are centered around zero), but that the predictions can be as much as 1,000 g off when using only `bill_length_mm` to predict `body_mass_g`. On the other hand, when using `bill_length_mm`, `bill_depth_mm`, `flipper_length_mm`, `sex`, and `species` to predict `body_mass_g`, the prediction errors seem to be about half as big, as seen in Figure 25.7. For the model using `bill_length_mm`, `bill_depth_mm`, `flipper_length_mm`, `sex`, and `species` to predict `body_mass_g`, the CV SSE is 27,728,698 (as compared to a CV SSE of 141,552,822 for the smaller model). Consistent with visually comparing the two sets of residual plots, the sum of squared prediction errors is smaller for the model which uses more predictor variables. The model with more predictor variables seems like the better model (according to the cross-validated prediction errors criteria).

We have provided a very brief overview to and example of using cross-validation. Cross-validation is a computational approach to model building and model validation as an alternative to reliance on p-values. While p-values have a role to play in understanding model coefficients, throughout this text, we have continued to present computational methods that broaden statistical approaches to data analysis. Cross-validation will be used again in Chapter 26 with logistic regression. We encourage you to consider both standard inferential methods (such as p-values) and computational approaches (such as cross-validation) as you build and use multivariable models of all varieties.

25.4 Chapter review

25.4.1 Summary

Building on the modeling ideas from Chapter 8, we have now introduced methods for evaluating coefficients (based on p-values) and evaluating models (cross-validation). There are many important aspects to consider when working with multiple variables in a single model, and we have only glanced at a few topics. Remember, multicollinearity can make coefficient interpretation difficult. A topic not covered in this text but important for multiple regression models is interaction, and we hope that you learn more about how variables work together as you continue to build up your modeling skills.

25.4.2 Terms

The terms introduced in this chapter are presented in Table 25.7. If you're not sure what some of these terms mean, we recommend you go back in the text and review their definitions. You should be able to easily spot them as **bolded text**.

Table 25.7: Terms introduced in this chapter.

cross-validation	multicollinearity	prediction error
inference on multiple linear regression	multiple predictors	predictor

25.5 Exercises

Answers to odd-numbered exercises can be found in Appendix A.25.

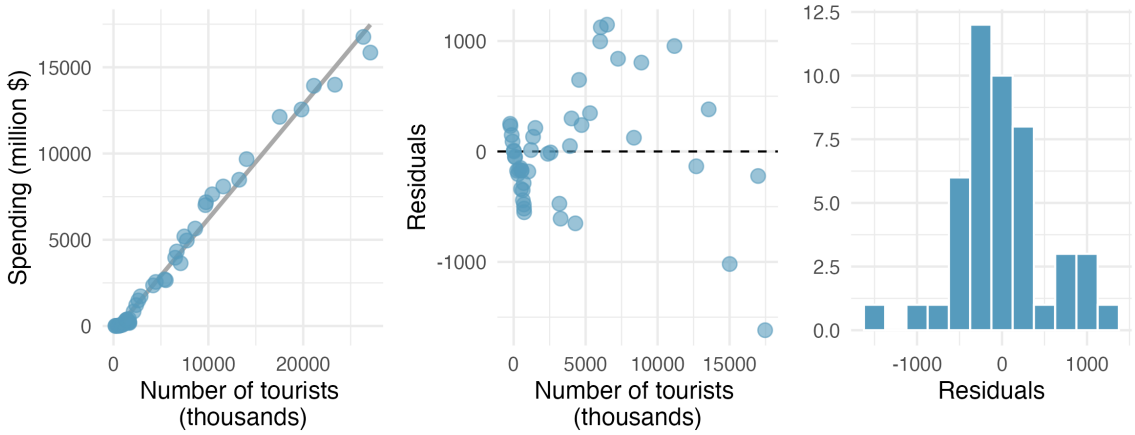
1. **GPA, mathematical interval.** A survey of 55 Duke University students asked about their GPA (`gpa`), number of hours they study weekly (`studyweek`), number of hours they sleep nightly (`sleepnight`), and whether they go out more than two nights a week (`out_mt2`). We use these data to build a model predicting GPA from the other variables. Summary of the model is shown below. Note that `out_mt2` is `1` if the student goes out more than two nights a week, and 0 otherwise.[3]

term	estimate	std.error	statistic	p.value
`(Intercept)`	3.508	0.347	10.114	<0.0001
`studyweek`	0.002	0.004	0.400	0.6908
`sleepnight`	0.000	0.047	0.008	0.994
`out_mt2`	0.151	0.097	1.551	0.127

   a. Calculate a 95% confidence interval for the coefficient of `out_mt2` (go out more than two night a week) in the model, and interpret it in the context of the data.

   b. Would you expect a 95% confidence interval for the slope of the remaining variables to include 0? Explain your reasoning.

2. **Tourism spending.** The Association of Turkish Travel Agencies reports the number of foreign tourists visiting Turkey and tourist spending by year. Three plots are provided: scatterplot showing the relationship between these two variables along with the least squares fit, residuals plot, and histogram of residuals.[4]



   a. Describe the relationship between number of tourists and spending.

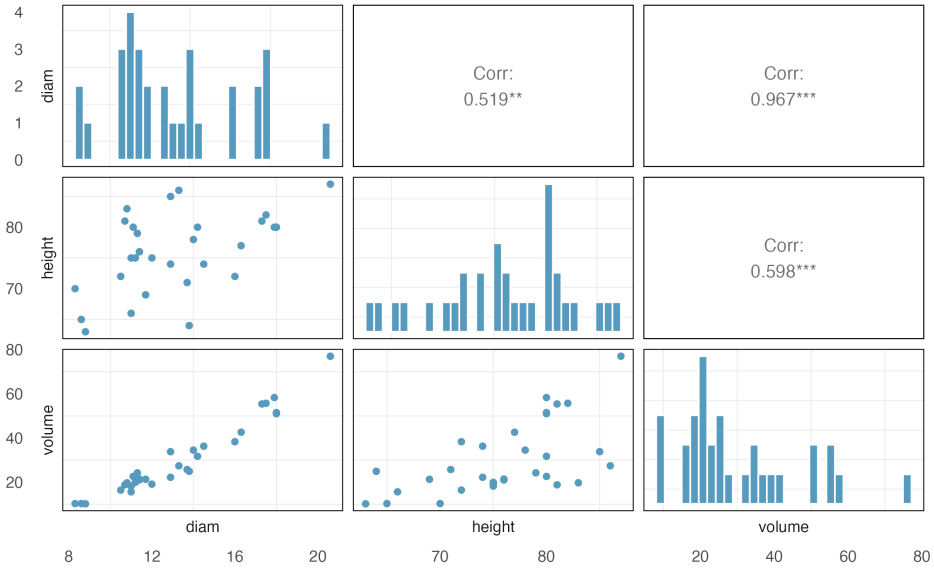
   b. What are the predictor and the outcome variables?

   c. Why might we want to fit a regression line to these data?

   d. Do the data meet the LINE conditions required for fitting a least squares line? Use the scatterplot, the residuals plot, and the histogram to answer this question.

[3] The `gpa` data used in this exercise can be found in the **openintro** R package.
[4] The `tourism` data used in this exercise can be found in the **openintro** R package.

3. **Cherry trees, collinear predictors.** Timber yield is approximately equal to the volume of a tree, however, this value is difficult to measure without first cutting the tree down. Instead, other variables, such as height and diameter, may be used to predict a tree's volume and yield. Researchers wanting to understand the relationship between these variables for black cherry trees collected data from 31 such trees in the Allegheny National Forest, Pennsylvania. Height is measured in feet, diameter in inches (at 54 inches above ground), and volume in cubic feet.[5] (Hand 1994) The plots below display the distribution of each of these variables (on the diagonal) as well as provide information on the pairwise correlations between them.



Also provided below are three regression model outputs: `volume` vs. `diam`, `volume` vs. `height`, and `volume` vs. `height + diam`.

term	estimate	std.error	statistic	p.value
`(Intercept)`	-36.94	3.365	-11.0	<0.0001
`diam`	5.07	0.247	20.5	<0.0001

term	estimate	std.error	statistic	p.value
`(Intercept)`	-87.12	29.273	-2.98	0.006
`height`	1.54	0.384	4.02	0.000

term	estimate	std.error	statistic	p.value
`(Intercept)`	-57.988	8.638	-6.71	<0.0001
`height`	0.339	0.130	2.61	0.0145
`diam`	4.708	0.264	17.82	<0.0001

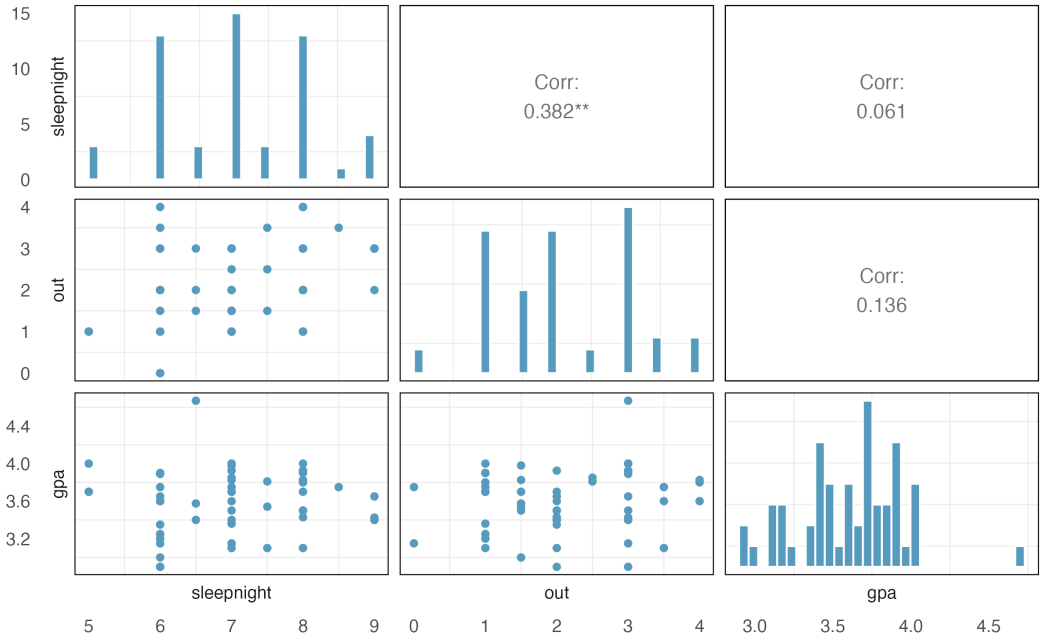
a. There are three variables described in the figure, and each is paired with each other to create three different scatterplots. Rate the pairwise relationships from most correlated to least correlated.

b. When using only one variable to model a tree's `volume`, is `diam`eter a discernible predictor? Is `height` a discernible predictor? Explain your reasoning.

c. When using both `diam`eter and `height` to predict a tree's `volume`, are both predictors still discernible? Explain your reasoning.

[5]The `cherry` data used in this exercise can be found in the openintro R package.

4. **GPA, collinear predictors.** In this exercise we work with data from a survey of 55 Duke University students who were asked about their GPA, number of hours they sleep nightly, and number of nights they go `out` each week. The plots below display the distributions of each of these variables (on the diagonal) as well as their pairwise relationships and correlation coefficients.



Also provided below are three regression model outputs: `gpa` vs. `out`, `gpa` vs. `sleepnight`, and `gpa` vs. `out + sleepnight`.

term	estimate	std.error	statistic	p.value
`(Intercept)`	3.504	0.106	33.011	<0.0001
`out`	0.045	0.046	0.998	0.3229

term	estimate	std.error	statistic	p.value
`(Intercept)`	3.46	0.318	10.874	<0.0001
`sleepnight`	0.02	0.045	0.445	0.6583

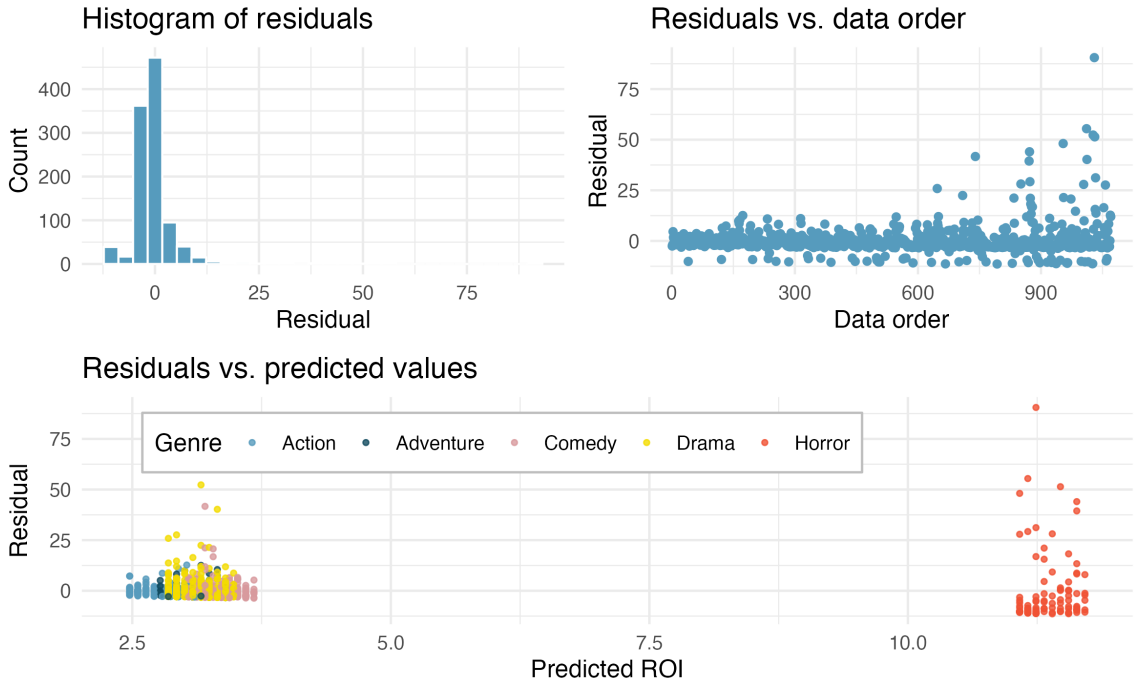
term	estimate	std.error	statistic	p.value
`(Intercept)`	3.483	0.320	10.888	<0.0001
`out`	0.044	0.050	0.886	0.3796
`sleepnight`	0.003	0.048	0.072	0.9432

a. There are three variables described in the figure, and each is paired with each other to create three different scatterplots. Rate the pairwise relationships from most correlated to least correlated.

b. When using only one variable to model `gpa`, is `out` a discernible predictor? Is `sleepnight` a discernible predictor? Explain your reasoning.

c. When using both `out` and `sleepnight` to predict `gpa` in a multiple regression model, are either of the variables discernible? Explain your reasoning.

5. **Movie returns.** A FiveThirtyEight.com article reports that "Horror movies get nowhere near as much draw at the box office as the big-time summer blockbusters or action/adventure movies, but there's a huge incentive for studios to continue pushing them out. The return-on-investment potential for horror movies is absurd." To investigate how the return-on-investment (ROI) compares between genres and how this relationship has changed over time, an introductory statistics student fit a linear regression model to predict the ratio of gross revenue of movies to the production costs from genre and release year for 1,070 movies released between 2000 and 2018. Using the plots given below, determine if this regression model is appropriate for these data. In particular, use the residual plot to check the LINE conditons. (FiveThirtyEight 2015)



6. **Difficult encounters.** A study was conducted at a university outpatient primary care clinic in Switzerland to identify factors associated with difficult doctor-patient encounters. The data consist of 527 patient encounters, conducted by the 27 medical residents employed at the clinic. After each encounter, the attending physician completed two questionnaires: the Difficult Doctor Patient Relationship Questionnaire (DDPRQ-10) and the Patient's Vulnerability Grid (PVG). A higher score on the DDPRQ-10 indicates a more difficult encounter. The maximum possible score is 60 and encounters with score 30 and higher are considered difficult. A model was fit to predict DDPRQ-10 score from features of the attending physician: age, sex (male or not), and years of training.

term	estimate	std.error	statistic	p.value
`(Intercept)`	30.594	2.886	10.601	<0.0001
`age`	-0.016	0.104	-0.157	0.876
`sexMale`	-0.535	0.781	-0.686	0.494
`yrs_train`	0.096	0.215	0.445	0.656

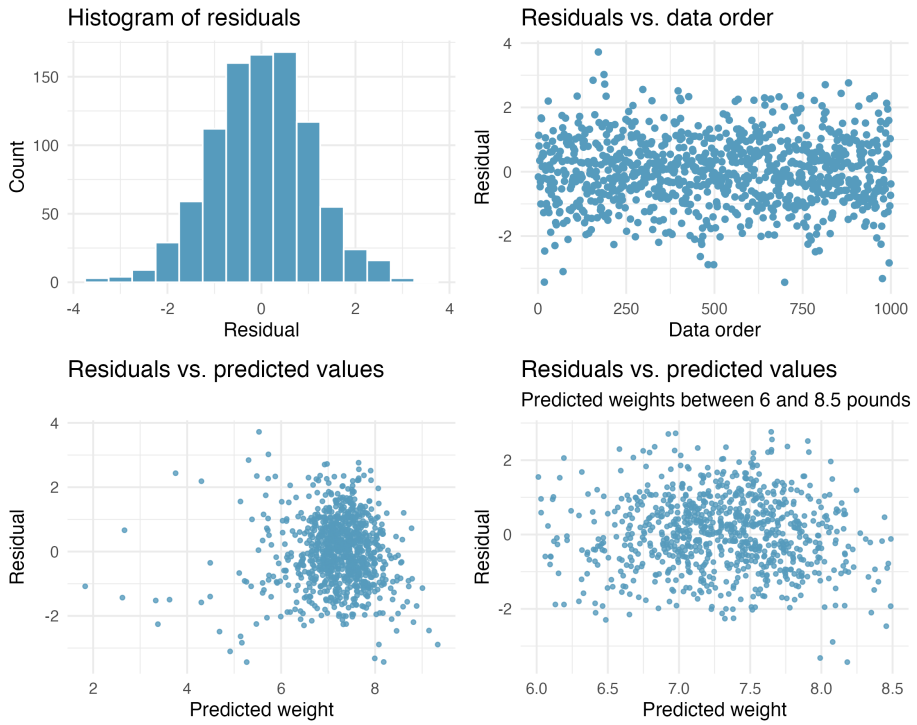
   a. The intercept of the model is 30.594. What is the age, sex, and years of training of a physician whom this model would predict to have a DDPRQ-10 score of 30.594.

   b. Is there evidence of a discernible association between DDPRQ-10 score and any of the physician features?

7. **Baby's weight, mathematical test.** US Department of Health and Human Services, Centers for Disease Control and Prevention collect information on births recorded in the country. The data used here are a random sample of 1,000 births from 2014. Here, we study the relationship between smoking and weight of the baby. The variable `smoke` is coded 1 if the mother is a smoker, and 0 if not. The summary table below shows the results of a linear regression model for predicting the average birth weight of babies, measured in pounds, based on the smoking status of the mother.[6] (ICPSR 2014)

term	estimate	std.error	statistic	p.value
`(Intercept)`	-3.82	0.57	-6.73	<0.0001
`weeks`	0.26	0.01	18.93	<0.0001
`mage`	0.02	0.01	2.53	0.0115
`sexmale`	0.37	0.07	5.30	<0.0001
`visits`	0.02	0.01	2.09	0.0373
`habitsmoker`	-0.43	0.13	-3.41	7e-04

Also shown below are a series of diagnostics plots.

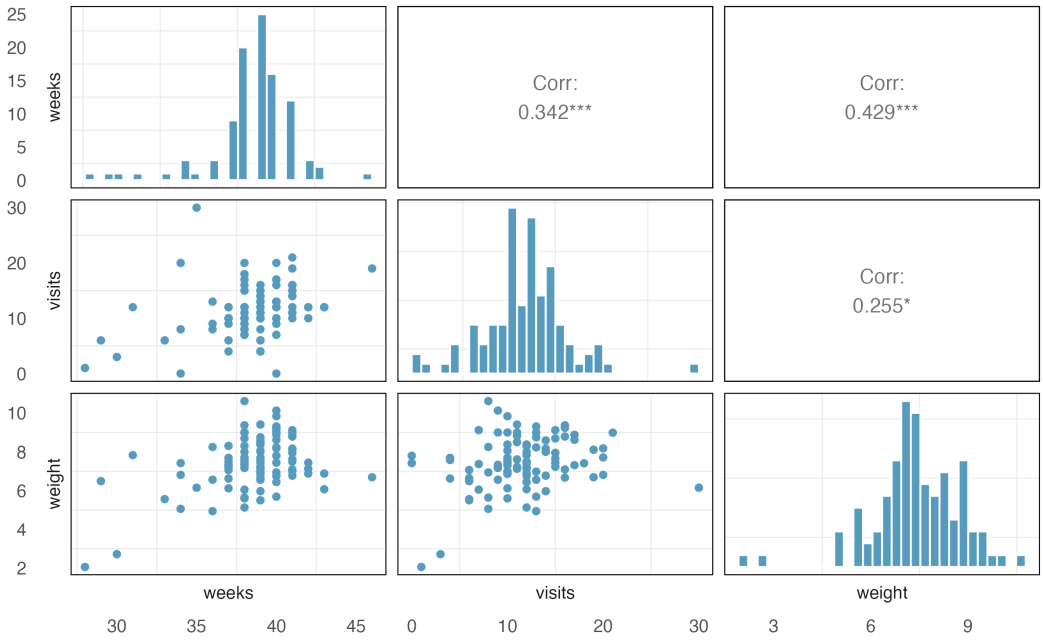


a. Determine if the conditions for doing inference based on mathematical models with these data are met using the diagnostic plots above. If not, describe how to proceed with the analysis.

b. Using the regression output, evaluate whether the true slope of `habit` (i.e., whether the mother is a smoker) is different than 0, given the other variables in the model. State the null and alternative hypotheses, report the p-value (using a mathematical model), and state your conclusion.

[6] The `births14` data used in this exercise can be found in the openintro R package.

8. **Baby's weight, collinear predictors.** In this exercise we study the relationship between the weight of the baby and two explanatory variables: number of `weeks` of gestation and number of pregnancy hospital `visits`. (ICPSR 2014) The plots below display the distributions of each of these variables (on the diagonal) as well as their pairwise relationships and correlation coefficients.



Also provided below are three regression model outputs: `weight` vs. `weeks`, `weight` vs. `visits`, and `weight` vs. `weeks + visits`.

term	estimate	std.error	statistic	p.value
`(Intercept)`	-0.83	1.76	-0.47	0.6395
`weeks`	0.21	0.05	4.70	<0.0001

term	estimate	std.error	statistic	p.value
`(Intercept)`	6.56	0.36	18.36	<0.0001
`visits`	0.08	0.03	2.61	0.0105

term	estimate	std.error	statistic	p.value
`(Intercept)`	-0.44	1.78	-0.25	0.81
`weeks`	0.19	0.05	4.00	0.00
`visits`	0.04	0.03	1.26	0.21

a. There are three variables described in the figure, and each is paired with each other to create three different scatterplots. Rate the pairwise relationships from most correlated to least correlated.

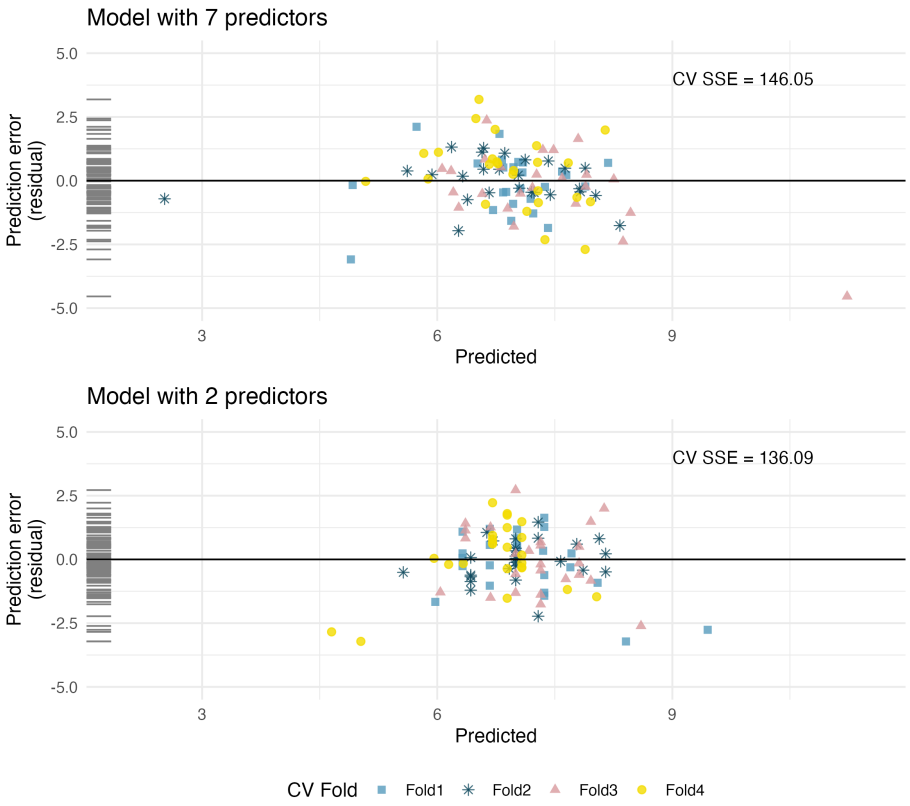
b. When using only one variable to model the baby's `weight`, is `weeks` a discernible predictor? Is `visits` a discernible predictor? Explain your reasoning.

c. When using both `visits` and `weeks` to predict the baby's `weight`, are both predictors still discernible? Explain your reasoning.

9. **Baby's weight, cross-validation.** Using a random sample of 1,000 US births from 2014, we study the relationship between the weight of the baby and various explanatory variables. (ICPSR 2014) The plots below display prediction errors associated with two different models designed to predict `weight` of baby at birth; one model uses 7 predictors, one model uses 2 predictors. Using 4-fold cross-validation, the data were split into 4 folds. Three of the folds estimate the β_i parameters using b_i, and the model is applied to the held out fold for prediction. The process was repeated 4 times (each time holding out one of the folds).

$$E[\texttt{weight}] = \beta_0 + \beta_1 \times \texttt{fage} + \beta_2 \times \texttt{mage} + \beta_3 \times \texttt{mature} + \beta_4 \times \texttt{weeks} + \beta_5 \times \texttt{visits} + \beta_6 \times \texttt{gained} + \beta_7 \times \texttt{habit}$$

$$E[\texttt{weight}] = \beta_0 + \beta_1 \times \texttt{weeks} + \beta_2 \times \texttt{mature}$$



a. In the first graph, note the point at roughly (predicted = 11 and error = -4). Estimate the observed and predcited value for that observation.

b. Using the same point, describe which cross-validation fold(s) were used to build its prediction model.

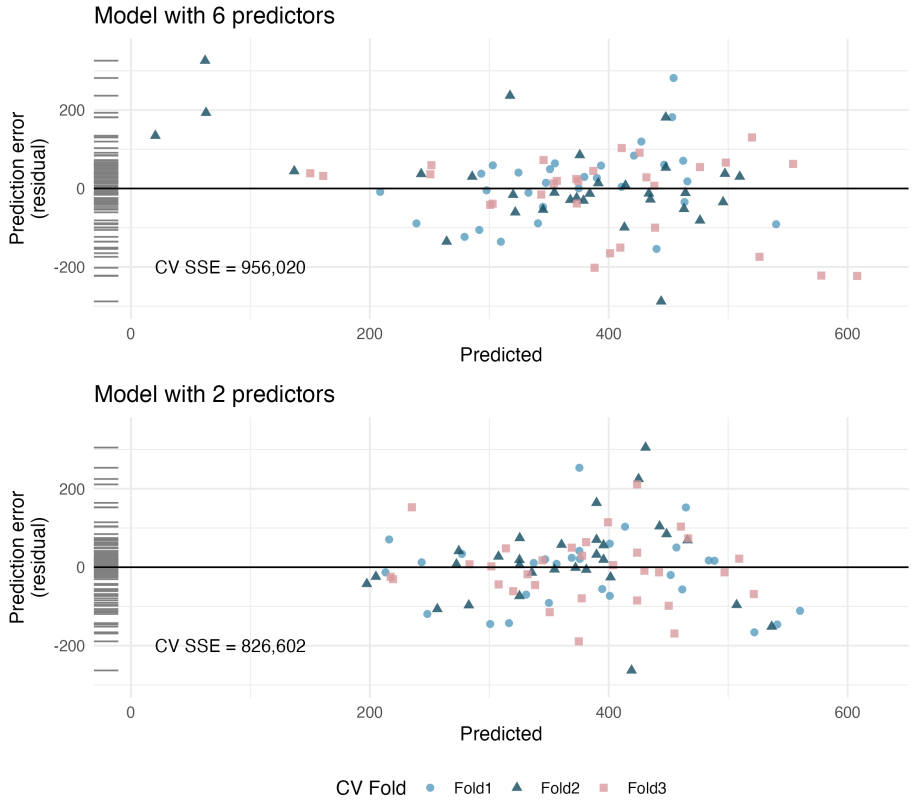
c. For the plot on the top, for one of the cross-validation folds, how many coefficients were estimated in the linear model? For the plot on the bottom, for one of the cross-validation folds, how many coefficients were estimated in the linear model?

d. Do the values of the residuals (along the y-axis, not the x-axis) seem markedly different for the two models? Explain your reasoning.

10. **RailTrail, cross-validation.** The Pioneer Valley Planning Commission (PVPC) collected data north of Chestnut Street in Florence, MA for ninety days from April 5, 2005 to November 15, 2005. Data collectors set up a laser sensor, with breaks in the laser beam recording when a rail-trail user passed the data collection station.[7] The plots below display prediction errors associated with two different models designed to predict the `volume` of riders on the RailTrail; one model uses 6 predictors, one model uses 2 predictors. Using 3-fold cross-validation, the data were split into 3 folds. Three of the folds estimate the β_i parameters using b_i, and the model is applied to the held out fold for prediction. The process was repeated 4 times (each time holding out one of the folds).

$$E[\texttt{volume}] = \beta_0 + \beta_1 \times \texttt{hightemp} + \beta_2 \times \texttt{lowtemp} + \beta_3 \times \texttt{spring} + \beta_4 \times \texttt{weekday} + \beta_5 \times \texttt{cloudcover} + \beta_6 \times \texttt{precip}$$

$$E[\texttt{volume}] = \beta_0 + \beta_1 \times \texttt{hightemp} + \beta_2 \times \texttt{precip}$$



a. In the second graph, note the point at roughly (predicted = 400 and error = 100). Estimate the observed and predcited value for that observation.

b. Using the same point, describe which cross-validation fold(s) were used to build its prediction model.

c. For the plot on the top, for one of the cross-validation folds, how many coefficients were estimated in the linear model? For the plot on the bottom, for one of the cross-validation folds, how many coefficients were estimated in the linear model?

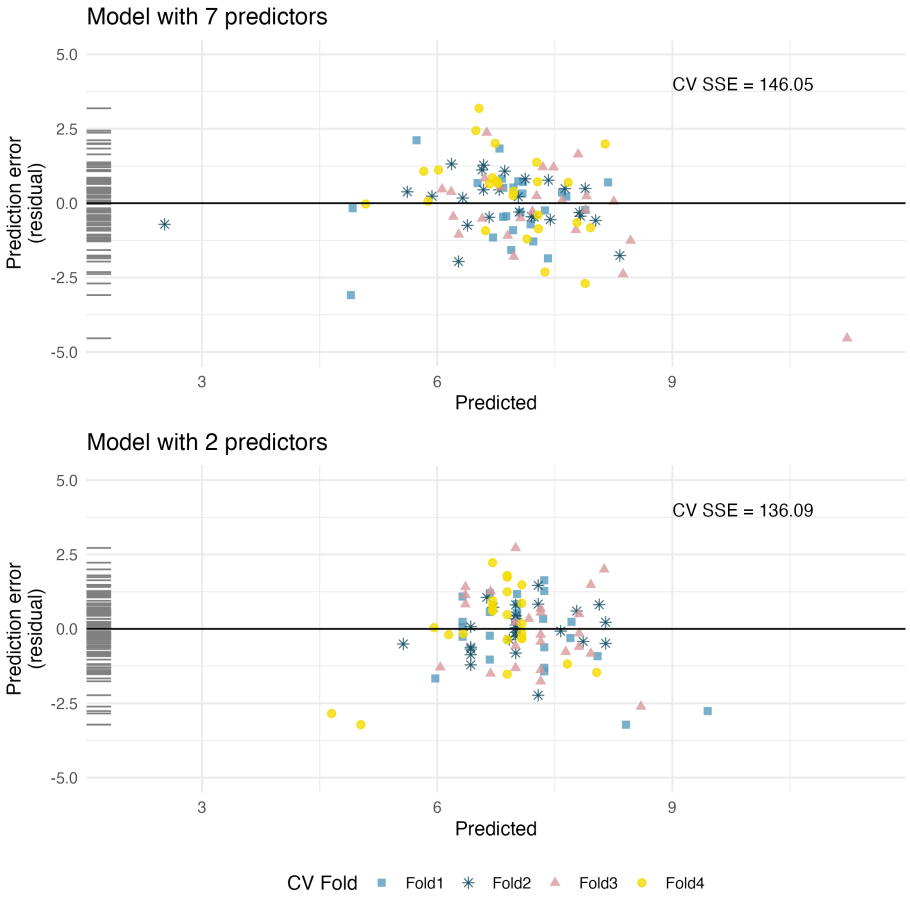
d. Do the values of the residuals (along the y-axis, not the x-axis) seem markedly different for the two models? Explain your reasoning.

[7] The `RailTrail` data used in this exercise can be found in the **mosaicData** R package.

11. **Baby's weight, cross-validation for model selection.** Using a random sample of 1,000 US births from 2014, we study the relationship between the weight of the baby and various explanatory variables. (ICPSR 2014) The plots below display prediction errors associated with two different models designed to predict `weight` of baby at birth; one model uses 7 predictors, one model uses 2 predictors. Using 4-fold cross-validation, the data were split into 4 folds. Three of the folds estimate the β_i parameters using b_i, and the model is applied to the held out fold for prediction. The process was repeated 4 times, each time holding out one of the folds.

$$\begin{aligned} E[\texttt{weight}] = \beta_0 &+ \beta_1 \times \texttt{fage} + \beta_2 \times \texttt{mage} \\ &+ \beta_3 \times \texttt{mature} + \beta_4 \times \texttt{weeks} \\ &+ \beta_5 \times \texttt{visits} + \beta_6 \times \texttt{gained} \\ &+ \beta_7 \times \texttt{habit} \end{aligned}$$

$$E[\texttt{weight}] = \beta_0 + \beta_1 \times \texttt{weeks} + \beta_2 \times \texttt{mature}$$



a. Using the spread of the points (in the y-direction), which model should be chosen for a final report on these data? Explain your reasoning.

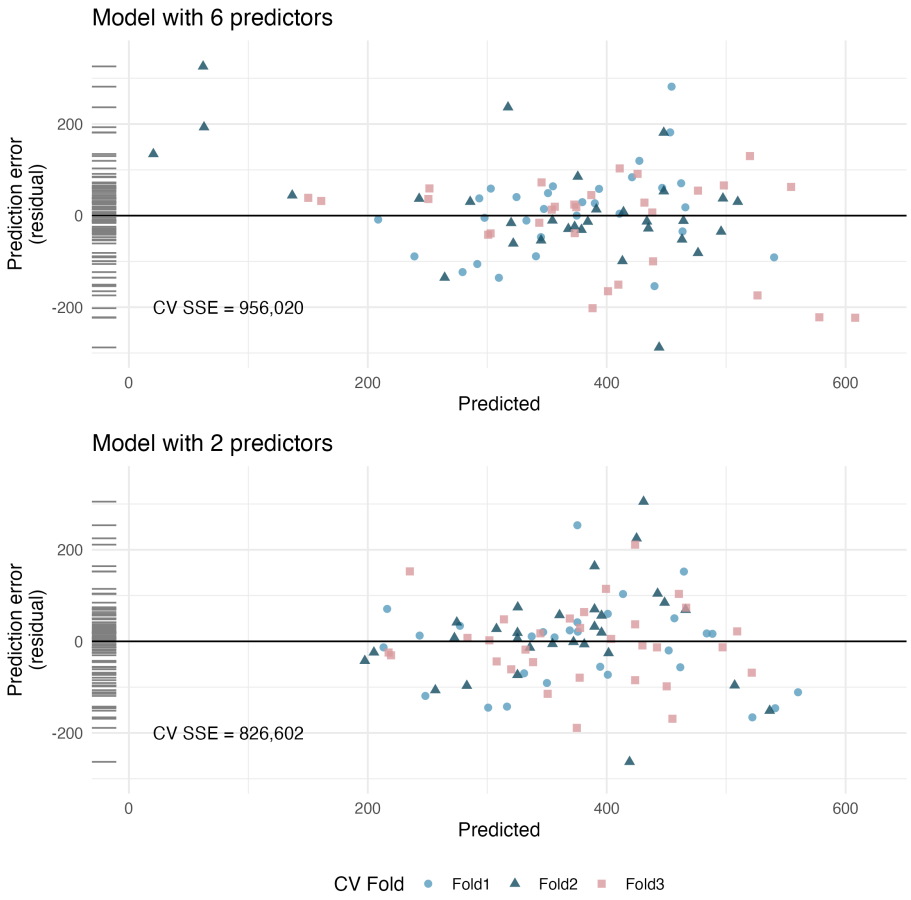
b. Using the summary statistic (CV SSE), which model should be chosen for a final report on these data? Explain your reasoning.

c. Why would the model with more predictors fit the data less closely than the model with only two predictors?

12. **RailTrail, cross-validation for model selection.** The Pioneer Valley Planning Commission (PVPC) collected data north of Chestnut Street in Florence, MA for ninety days from April 5, 2005 to November 15, 2005. Data collectors set up a laser sensor, with breaks in the laser beam recording when a rail-trail user passed the data collection station. The plots below display prediction errors associated with two different models designed to predict the `volume` of riders on the RailTrail; one model uses 6 predictors, one model uses 2 predictors. Using 3-fold cross-validation, the data were split into 3 folds. Three of the folds estimate the β_i parameters using b_i, and the model is applied to the held out fold for prediction. The process was repeated 4 times, each time holding out one of the folds.

$$E[\texttt{volume}] = \beta_0 + \beta_1 \times \texttt{hightemp} + \beta_2 \times \texttt{lowtemp} + \beta_3 \times \texttt{spring} + \beta_4 \times \texttt{weekday} + \beta_5 \times \texttt{cloudcover} + \beta_6 \times \texttt{precip}$$

$$E[\texttt{volume}] = \beta_0 + \beta_1 \times \texttt{hightemp} + \beta_2 \times \texttt{precip}$$



a. Using the spread of the points (in the y-direction), which model should be chosen for a final report on these data? Explain your reasoning.

b. Using the summary statistic (CV SSE), which model should be chosen for a final report on these data? Explain your reasoning.

c. Why would the model with more predictors fit the data less closely than the model with only two predictors?

Chapter 26

Inference for logistic regression

Combining ideas from Chapter 9 on logistic regression, Chapter 13 on inference with mathematical models, and Chapter 24 and Chapter 25 which apply inferential techniques to the linear model, we wrap up the book by considering inferential methods applied to a logistic regression model. Additionally, we use cross-validation as a method for independent assessment of the logistic regression model.

As with multiple linear regression, the inference aspect for logistic regression will focus on interpretation of coefficients and relationships between explanatory variables. Both p-values and cross-validation will be used for assessing a logistic regression model.

Consider the `email` data which describes email characteristics which can be used to predict whether a particular incoming email is spam (unsolicited bulk email). Without reading every incoming message, it might be nice to have an automated way to identify spam emails. Which of the variables describing each email are important for predicting the status of the email?

The `email` data can be found in the **openintro** R package.

26.1 Model diagnostics

Before looking at the hypothesis tests associated with the coefficients (turns out they are very similar to those in linear regression!), it is valuable to understand the technical conditions that underlie the inference applied to the logistic regression model. Generally, as you've seen in the logistic regression modeling examples, it is imperative that the response variable is binary. Additionally, the key technical condition for logistic regression has to do with the relationship between the predictor variables (x_i values) and the probability the outcome will be a success. It turns out, the relationship is a specific functional form called a logit function, where $\text{logit}(p) = \log_e(\frac{p}{1-p})$. The function may feel complicated, and memorizing the formula of the logit is not necessary for understanding logistic regression. What you do need to remember is that the probability of the outcome being a success is a function of a linear combination of the explanatory variables.

Table 26.1: Variables and their descriptions for the `email` dataset. Many of the variables are indicator variables, meaning they take the value 1 if the specified characteristic is present and 0 otherwise.

Variable	Description
`spam`	Indicator for whether the email was spam.
`to_multiple`	Indicator for whether the email was addressed to more than one recipient.
`from`	Whether the message was listed as from anyone (this is usually set by default for regular outgoing email).
`cc`	Number of people cc'ed.
`sent_email`	Indicator for whether the sender had been sent an email in the last 30 days.
`attach`	The number of attached files.
`dollar`	The number of times a dollar sign or the word "dollar" appeared in the email.
`winner`	Indicates whether "winner" appeared in the email.
`format`	Indicates whether the email was written using HTML (e.g., may have included bolding or active links).
`re_subj`	Whether the subject started with "Re:", "RE:", "re:", or "rE:"
`exclaim_subj`	Whether there was an exclamation point in the subject.
`urgent_subj`	Whether the word "urgent" was in the email subject.
`exclaim_mess`	The number of exclamation points in the email message.
`number`	Factor variable saying whether there was no number, a small number (under 1 million), or a big number.

Logistic regression conditions.

There are two key conditions for fitting a logistic regression model:

1. Each outcome Y_i is independent of the other outcomes.
2. Each predictor x_i is linearly related to $\text{logit}(p_i)$ if all other predictors are held constant.

The first logistic regression model condition — independence of the outcomes — is reasonable if we can assume that the emails that arrive in an inbox within a few months are independent of each other with respect to whether they're spam or not.

The second condition of the logistic regression model is not easily checked without a fairly sizable amount of data. Luckily, we have 3921 emails in the dataset! Let's first visualize these data by plotting the true classification of the emails against the model's fitted probabilities, as shown in Figure 26.1.

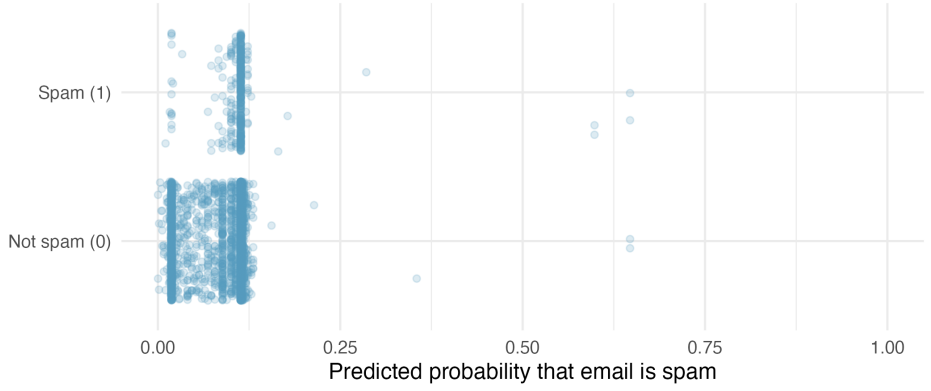


Figure 26.1: The predicted probability that each of the 3921 emails are spam. Points have been jittered so that those with nearly identical values aren't plotted exactly on top of one another.

We'd like to assess the quality of the model. For example, we might ask: if we look at emails that we modeled as having 10% chance of being spam, do we find out 10% of the actually are spam? We can check this for groups of the data by constructing a plot as follows:

1. Bucket the observations into groups based on their predicted probabilities.
2. Compute the average predicted probability for each group.
3. Compute the observed probability for each group, along with a 95% confidence interval for the true probability of success for those individuals.
4. Plot the observed probabilities (with 95% confidence intervals) against the average predicted probabilities for each group.

If the model does a good job describing the data, the plotted points should fall close to the line $y = x$, since the predicted probabilities should be similar to the observed probabilities. We can use the confidence intervals to roughly gauge whether anything might be amiss. Such a plot is shown in Figure 26.2.

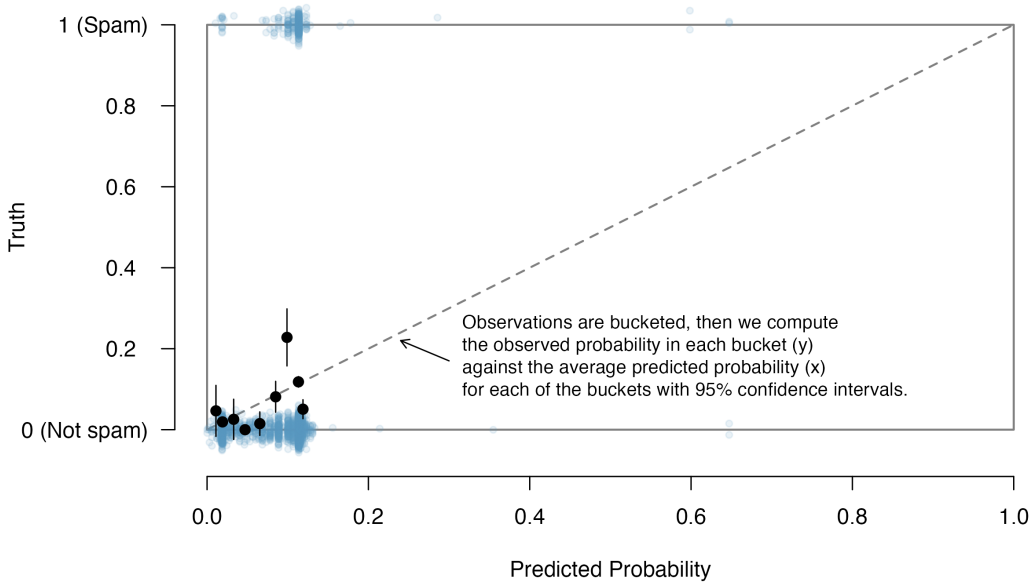


Figure 26.2: A reconfiguration of Figure 26.1. Again, the predicted probabilities are on the x-axis and the truth is on the y-axis for each email. After data have been bucketed into predicted probability groups, the proportion of spam emails (i.e., the observed probability) is given by the black circles. The dashed line is within the confidence bound of the 95% confidence intervals for many of the buckets, suggesting the logistic fit is reasonable.

A plot like Figure 26.2 helps to better understand the deviations. Additional diagnostics may be created that are similar to those featured in Section 24.6. For instance, we could compute residuals as the observed outcome minus the expected outcome ($e_i = Y_i - \hat{p}_i$), and then we could create plots of these residuals against each predictor.

26.2 Multiple logistic regression output from software

As you learned in Chapter 8, optimization can be used to find the coefficient estimates for the logistic model. The unknown population model can be written as:

$$\log_e\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \times \texttt{to\_multiple} + \beta_2 \times \texttt{cc} + \beta_3 \times \texttt{dollar} + \beta_4 \times \texttt{urgent\_subj}$$

The estimated equation for the regression model may be written as a model with four predictor variables, where $\hat{p}$ is the estimated probability of being a spam email message:

$$\begin{aligned}\log_e\left(\frac{\hat{p}}{1-\hat{p}}\right) = &-2.05 - 1.91 \times \texttt{to\_multiple} \\ &+ 0.02 \times \texttt{cc} \\ &- 0.07 \times \texttt{dollar} \\ &+ 2.66 \times \texttt{urgent\_subj}\end{aligned}$$

Table 26.2: Summary of a logistic model for predicting whether an email is spam based on the variables to_multiple, cc, dollar, and urgent_subj. Each of the variables has its own coefficient estimate and p-value.

term	estimate	std.error	statistic	p.value
(Intercept)	-2.05	0.06	-34.67	<0.0001
to_multiple1	-1.91	0.30	-6.37	<0.0001
cc	0.02	0.02	1.16	0.245
dollar	-0.07	0.02	-3.38	7e-04
urgent_subj1	2.66	0.80	3.32	9e-04

Not only does Table 26.2 provide the estimates for the coefficients, it also provides information on the inference analysis (i.e., hypothesis testing) which is the focus of this chapter.

As in Chapter 25, with **multiple predictors**, each hypothesis test (for each of the explanatory variables) is conditioned on each of the other variables remaining in the model.

if multiple predictors $H_0 : \beta_i = 0$ given other variables in the model

Using the example above and focusing on each of the variable p-values (here we won't discuss the p-value associated with the intercept), we can write out the four different hypotheses (associated with the p-value corresponding to each of the coefficients / rows in Table 26.2):

- $H_0 : \beta_1 = 0$ given `cc`, `dollar`, and `urgent_subj` are included in the model
- $H_0 : \beta_2 = 0$ given `to_multiple`, `dollar`, and `urgent_subj` are included in the model
- $H_0 : \beta_3 = 0$ given `to_multiple`, `cc`, and `urgent_subj` are included in the model
- $H_0 : \beta_4 = 0$ given `to_multiple`, `dollar`, and `dollar` are included in the model

The very low p-values from the software output tell us that three of the variables (that is, not `cc`) act as statistically discernible predictors in the model at the discernibility level of 0.05, despite the inclusion of any of the other variables. Consider the p-value on $H_0 : \beta_1 = 0$. The low p-value says that it would be extremely unlikely to observe data that yield a coefficient on `to_multiple` at least as far from 0 as -1.91 (i.e., $|b_1| > 1.91$) if the true relationship between `to_multiple` and `spam` was non-existent (i.e., if $\beta_1 = 0$) **and** the model also included `cc` and `dollar` and `urgent_subj`. Note also that the coefficient on `dollar` has a small associated p-value, but the magnitude of the coefficient is also seemingly small (0.07). It turns out that in units of standard errors (0.02 here), 0.07 is actually quite far from zero, it's all about context! The p-values on the remaining variables are interpreted similarly. From the initial output (p-values) in Table 26.2, it seems as though `to_multiple`, `dollar`, and `urgent_subj` are important variables for modeling whether an email is `spam`. We remind you that although p-values provide some information about the importance of each of the predictors in the model, there are many, arguably more important, aspects to consider when choosing the best model.

As with linear regression (see Section 25.2), existence of predictors that are correlated with each other can affect both the coefficient estimates and the associated p-values. However, investigating multicollinearity in a logistic regression model is saved for a text which provides more detail about logistic regression. Next, as a model building alternative (or enhancement) to p-values, we revisit cross-validation within the context of predicting status for each of the individual emails.

26.3 Cross-validation for prediction error

The p-value is a probability measure under a setting of no relationship. That p-value provides information about the degree of the relationship (e.g., above we measure the relationship between `spam` and `to_multiple` using a p-value), but the p-value does not measure how well the model will predict the individual emails (e.g., the accuracy of the model). Depending on the goal of the research project, you might be inclined to focus on variable importance (through p-values) or you might be inclined to focus on prediction accuracy (through cross-validation).

Here we present a method for using **cross-validation accuracy** to determine which variables (if any) should be used in a model which predicts whether an email is spam. A full treatment of cross-validation and logistic regression models is beyond the scope of this text. Using k-fold cross-validation, we can build k different models which are used to predict the observations in each of the k holdout samples. As with linear regression (see Section 25.3), we compare a smaller logistic regression model to a larger logistic regression model. The smaller model uses only the `to_multiple` variable, see the complete dataset (not cross-validated) model output in Table 26.3. The logistic regression model can be written as, where $\hat{p}$ is the estimated probability of being a spam email message:

The smaller model:

$$\log_e\left(\frac{\hat{p}}{1-\hat{p}}\right) = -2.12 + -1.81 \times \texttt{to\_multiple}$$

Table 26.3: The smaller model. Summary of a logistic model for predicting whether an email is spam based on only the predictor variable to_multiple. Each of the variables has its own coefficient estimate and p-value.

term	estimate	std.error	statistic	p.value
`(Intercept)`	-2.12	0.06	-37.67	<0.0001
`to_multiple1`	-1.81	0.30	-6.09	<0.0001

For each cross-validated model, the coefficients change slightly, and the model is used to make independent predictions on the holdout sample. The model from the first cross-validation sample is given in Figure 26.3 and can be compared to the coefficients in Table 26.3.

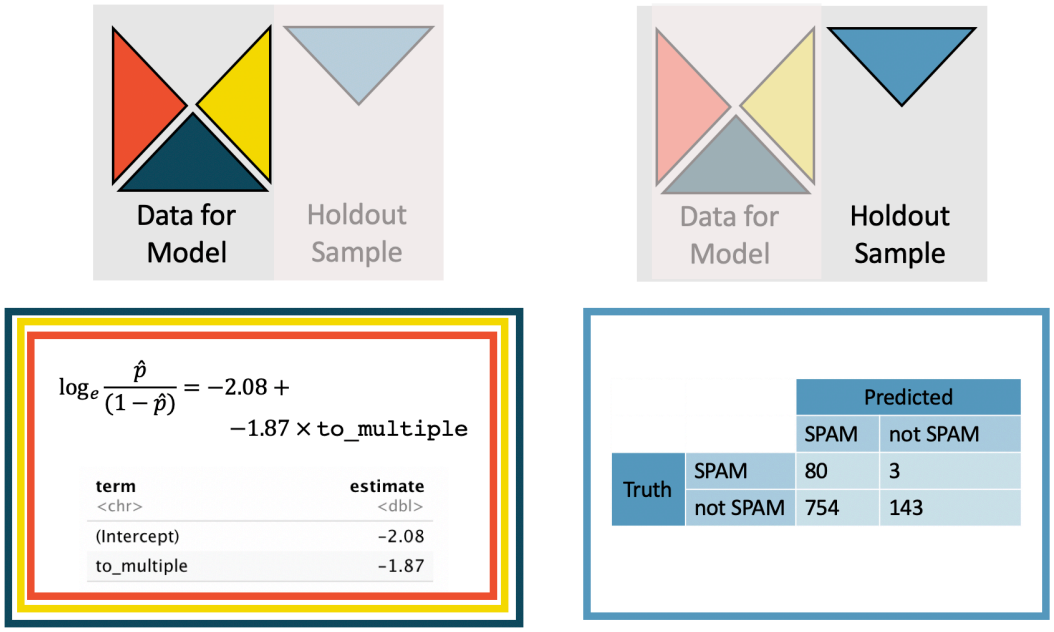


Figure 26.3: The smaller model. The coefficients are estimated using the least squares model on 3/4 of the data with a single predictor variable. Predictions are made on the remaining 1/4 of the observations. Note that the prediction error rate is quite high.

Table 26.4: The smaller model. One quarter at a time, the data were removed from the model building, and whether the email was spam (TRUE) or not (FALSE) was predicted. The logistic regression model was fit independently of the removed emails. Only `to_multiple` is used to predict whether the email is spam. Because we used a cutoff designed to identify spam emails, the accuracy of the non-spam email predictions is very low. `spamTP` is the proportion of true spam emails that were predicted to be spam. `notspamTP` is the proportion of true not spam emails that were predicted to be not spam.

fold	count	accuracy	notspamTP	spamTP
1st quarter	980	0.26	0.19	0.98
2nd quarter	981	0.23	0.15	0.96
3rd quarter	979	0.25	0.18	0.96
4th quarter	981	0.24	0.17	0.98

Because the `email` dataset has a ratio of roughly 90% non-spam and 10% spam emails, a model which randomly guessed all non-spam would have an overall accuracy of 90%! Clearly, we would like to capture the information with the spam emails, so our interest is in the percent of spam emails which are identified as spam (see Table 26.4). Additionally, in the logistic regression model, we use a 10% cutoff to predict whether the email is spam. Fortunately, we have done a great job of predicting! However, the trade-off was that most of the non-spam emails are now predicted to be spam which is not acceptable for a prediction algorithm. Adding more variables to the model may help with both the spam and non-spam predictions.

The larger model uses `to_multiple`, `attach`, `winner`, `format`, `re_subj`, `exclaim_mess`, and `number` as the set of seven predictor variables, see the complete dataset (not cross-validated) model output in Table 26.5. The logistic regression model can be written as follows, where $\hat{p}$ is the estimated probability of being a spam email message.

The larger model:

$$\begin{aligned}\log_e\left(\frac{\hat{p}}{1-\hat{p}}\right) = &-0.34 - 2.56 \times \texttt{to\_multiple} + 0.20 \times \texttt{attach} + 1.73 \times \texttt{winner}_{yes} \\ &- 1.28 \times \texttt{format} - 2.86 \times \texttt{re\_subj} + 0.00 \times \texttt{exclaim\_mess} \\ &- 1.07 \times \texttt{number}_{small} - 0.42 \times \texttt{number}_{big}\end{aligned}$$

Table 26.5: The larger model. Summary of a logistic model for predicting whether an email is spam based on the variables `to_multiple`, `attach`, `winner`, `format`, `re_subj`, `exclaim_mess`, and `number`. Each of the variables has its own coefficient estimate and p-value.

term	estimate	std.error	statistic	p.value
`(Intercept)`	-0.34	0.11	-3.02	0.0025
`to_multiple1`	-2.56	0.31	-8.28	<0.0001
`attach`	0.20	0.06	3.29	0.001
`winneryes`	1.73	0.33	5.33	<0.0001
`format1`	-1.28	0.13	-9.80	<0.0001
`re_subj1`	-2.86	0.37	-7.83	<0.0001
`exclaim_mess`	0.00	0.00	0.26	0.7925
`numbersmall`	-1.07	0.14	-7.54	<0.0001
`numberbig`	-0.42	0.20	-2.10	0.0357

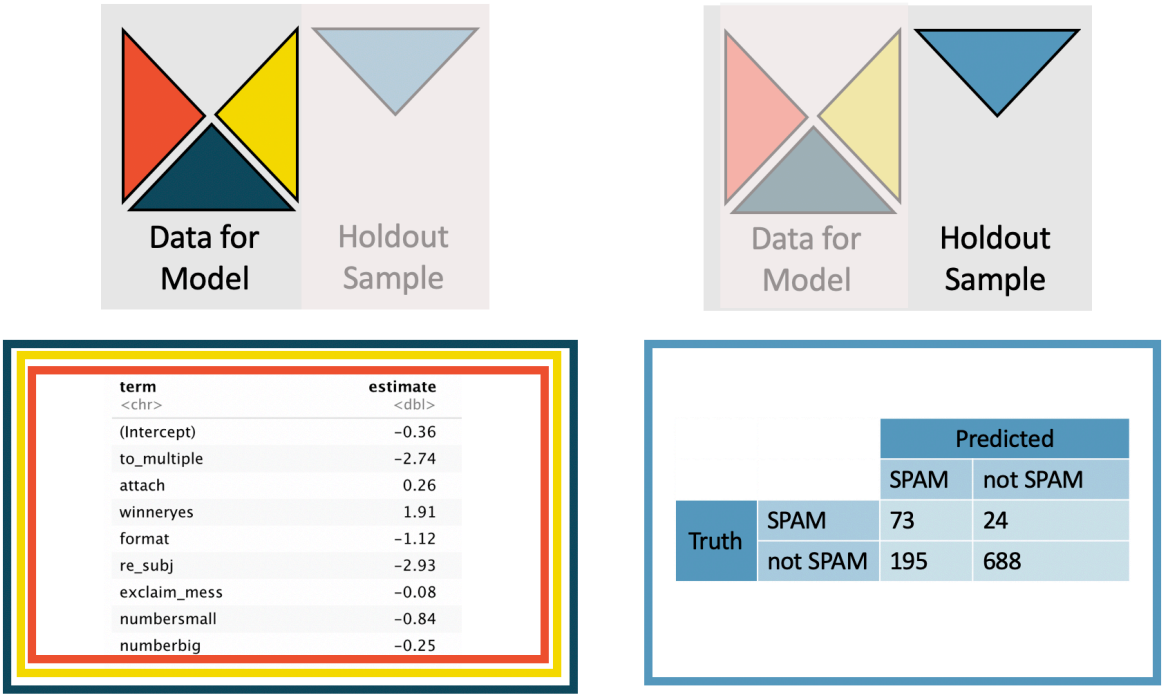


Figure 26.4: The larger model. The coefficients are estimated using the least squares model on 3/4 of the dataset with the seven specified predictor variables. Predictions are made on the remaining 1/4 of the observations. Note that the predictions are independent of the estimated model coefficients. The predictions are now much better for both the spam and the non-spam emails (than they were with a single predictor variable).

Table 26.6: The larger model. One quarter at a time, the data were removed from the model building, and whether the email was spam (TRUE) or not (FALSE) was predicted. The logistic regression model was fit independently of the removed emails. Now, the variables `to_multiple`, `attach`, `winner`, `format`, `re_subj`, `exclaim_mess`, and `number` are used to predict whether the email is spam. `spamTP` is the proportion of true spam emails that were predicted to be spam. `notspamTP` is the proportion of true not spam emails that were predicted to be not spam.'

fold	count	accuracy	notspamTP	spamTP
1st quarter	980	0.77	0.77	0.71
2nd quarter	981	0.80	0.81	0.70
3rd quarter	979	0.76	0.77	0.65
4th quarter	981	0.78	0.79	0.75

Somewhat expected, the larger model (see Table 26.6) was able to capture more nuance in the emails which lead to better predictions. However, it is not true that adding variables will always lead to better predictions, as correlated or noise variables may dampen the signal from the set of variables that truly predict the status. We encourage you to learn more about multiple variable models and cross-validation in your future exploration of statistical topics.

26.4 Chapter review

26.4.1 Summary

Throughout the text, we have presented a modern view to introduction to statistics. Earlier, we presented graphical techniques which communicated relationships across multiple variables. We also used modeling to formalize the relationships. In Chapter 26 we considered inferential claims on models which include many variables used to predict the probability of the outcome being a success. We continue to emphasize the importance of experimental design in making conclusions about research claims. In particular, recall that variability can come from different sources (e.g., random sampling vs. random allocation, see Figure 2.8).

As you might guess, this text has only scratched the surface of the world of statistical analyses that can be applied to different datasets. In particular, to do justice to the topic, the linear models and generalized linear models we have introduced can each be covered with their own course or book. Hierarchical models, alternative methods for fitting parameters (e.g., Ridge Regression or LASSO), and advanced computational methods applied to multivariable models (e.g., permuting the response variable? one explanatory variable? all the explanatory variables?) are all beyond the scope of this book. However, your successful understanding of the ideas we have covered has set you up perfectly to move on to a higher level of statistical modeling and inference. Enjoy!

26.4.2 Terms

The terms introduced in this chapter are presented in Table 26.7. If you're not sure what some of these terms mean, we recommend you go back in the text and review their definitions. You should be able to easily spot them as **bolded text**.

Table 26.7: Terms introduced in this chapter.

accuracy	inference on logistic regression	technical conditions
cross-validation	multiple predictors	

26.5 Exercises

Answers to odd-numbered exercises can be found in Appendix A.26.

1. **Passing the driver's test.** A consulting company is hired to assess whether certain characteristics of a DMV (e.g., location, number of test takers, number of test givers, hours of operation, etc.) are associated with the pass rate of the driver's test. The company is given information on 100 randomly selected DMVs across the country. For each DMV the annual driving test pass rate is recorded, along with other characteristics describing the DMV. Can logistic regression with multiple predictors be used to predict annual driving test pass rate for DMVs in this setting? Explain your reasoning.

2. **Oceans and skin cancer.** A researcher wants to investigate the relationship between living within 10 miles of an ocean for at least one year of life and developing skin cancer before the age of 50.

   a. Explain why logistic regression can be used to study the relationship between these two binary variables? What is the technical assumption describing the relationship between the response (outcome) and explanatory (predictor) variables?

   b. What other methods covered in this text might be used to address the research question of interest? What advantages does logistic regression have over these methods?

3. **Marijuana use in college.** Researchers studying whether the value systems of adolescents conflict with those of their parents asked 445 college students if they use marijuana. They also asked the students' parents if they used marijuana when they were in college. Based on the regression output shown below for predicting student drug use from parent drug use, evaluate whether parents' marijuana usage is a discernible predictor of their kids' marijuana usage. State the hypotheses, the test statistics, the p-value, and the conclusion in context of the data and the research question.[1] (Ellis and Stone 1979)

term	estimate	std.error	statistic	p.value
`(Intercept)`	-0.405	0.133	-3.04	0.0023
`parentsused`	0.791	0.194	4.09	<0.0001

4. **Treating heart attacks.** Researchers studying the effectiveness of Sulfinpyrazone in the prevention of sudden death after a heart attack conducted a randomized experiment on 1,475 patients. Based on the regression output shown below for predicting the outcome (`died` or `lived`, where success is defined as `lived`) from the treatment group (`control` and `treatment`), evaluate whether treatment group is a discernible predictor of the outcome. State the hypotheses, the test statistics, the p-value, and the conclusion in context of the data and the research question.[2] (Anturane Reinfarction Trial Research Group 1980)

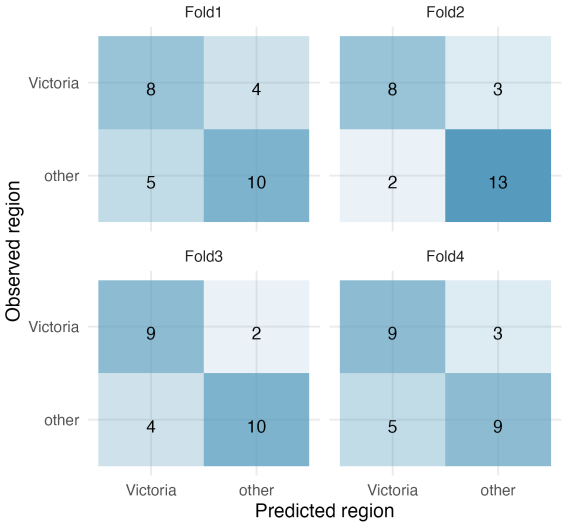
term	estimate	std.error	statistic	p.value
`(Intercept)`	2.431	0.135	18.05	<0.0001
`grouptreatment`	0.395	0.210	1.89	0.0594

[1] The `drug_use` data used in this exercise can be found in the openintro R package.
[2] The `sulphinpyrazone` data used in this exercise can be found in the openintro R package.

5. **Possum classification, cross-validation.** The common brushtail possum of the Australia region is a bit cuter than its distant cousin, the American opossum. We consider 104 brushtail possums from two regions in Australia, where the possums may be considered a random sample from the population. The first region is Victoria, which is in the eastern half of Australia and traverses the southern coast. The second region consists of New South Wales and Queensland, which make up eastern and northeastern Australia. We use logistic regression to differentiate between possums in these two regions. The outcome variable, called `pop`, takes value 1 when a possum is from Victoria and 0 when it is from New South Wales or Queensland.[3]

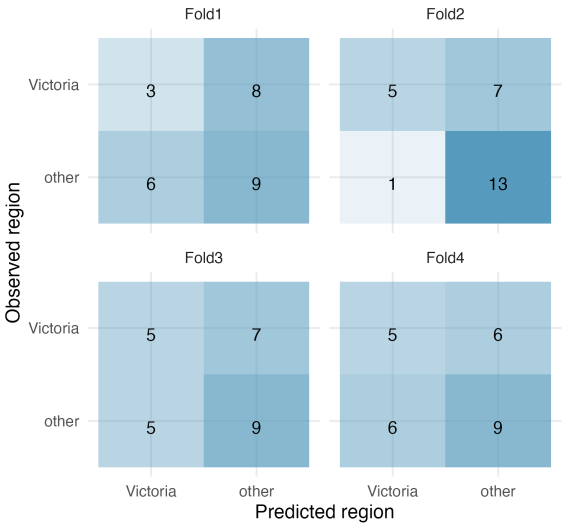
$$\log_e\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \times \texttt{tail\_l}$$

Model with tail length as predictor



$$\log_e\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \times \texttt{total\_l} + \beta_2 \times \texttt{sex}$$

Model with total length and sex as predictors



a. How many observations are in Fold2? Use the model with only tail length as a predictor variable. Of the observations in Fold2, how many of them were correctly predicted to be from Vicotria? How many of them were incorrectly predicted to be from Victoria?

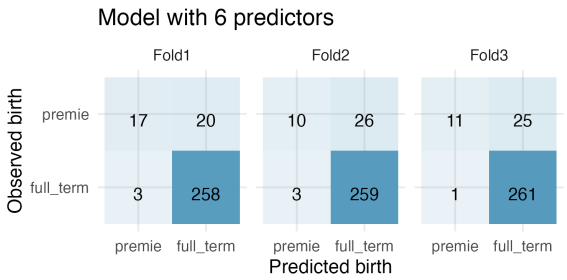
b. How many observations are used to build the model which predicts for the observations in Fold2?

c. For one of the cross-validation folds, how many coefficients were estimated for the model which uses tail length as a predictor? For one of the cross-validation folds, how many coefficients were estimated for the model which uses total length and sex as predictors?

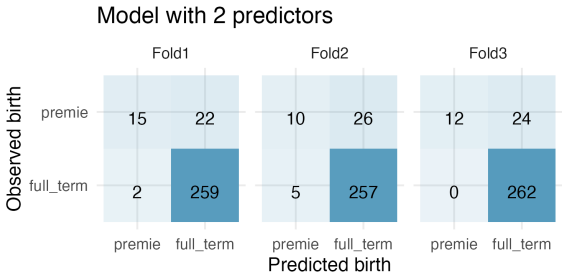
[3]The `possum` data used in this exercise can be found in the openintro R package.

6. **Premature babies, cross-validation.** US Department of Health and Human Services, Centers for Disease Control and Prevention collect information on births recorded in the country. The data used here are a random sample of 1000 births from 2014 (with some rows removed due to missing data). We use logistic regression to model whether the baby is premature from various explanatory variables.[4] (ICPSR 2014)

$$\log_e\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \times \texttt{mage} + \beta_2 \times \texttt{weight} + \beta_3 \times \texttt{mature} + \beta_4 \times \texttt{visits} + \beta_5 \times \texttt{gained} + \beta_6 \times \texttt{habit}$$



$$\log_e\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \times \texttt{weight} + \beta_2 \times \texttt{mature}$$



a. How many observations are in Fold2? Use the model with only `weight` and `mature` as predictor variables. Of the observations in Fold2, how many of them were correctly predicted to be premature? How many of them were incorrectly predicted to be premature?

b. How many observations are used to build the model which predicts for the observations in Fold2?

c. In the original dataset, are most of the births premature or full term? Explain.

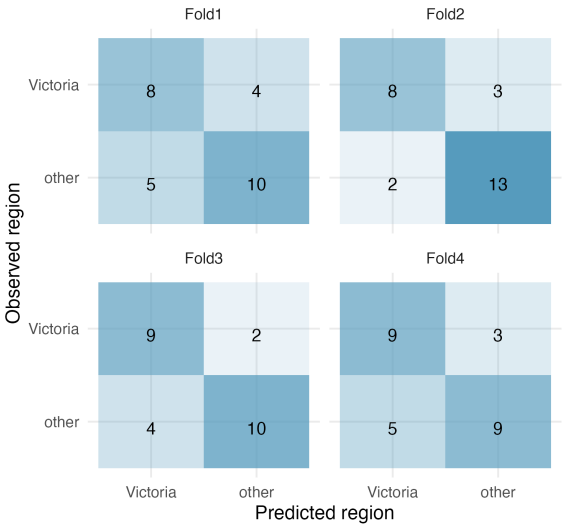
d. For one of the cross-validation folds, how many coefficients were estimated for the model which uses `mage`, `weight`, `mature`, `visits`, `gained`, and `habit` as predictors? For one of the cross-validation folds, how many coefficients were estimated for the model which uses `weight` and `mature` as predictors?

[4] The `births14` data used in this exercise can be found in the openintro R package.

7. **Possum classification, cross-validation to choose model.** In this exercise we consider 104 brushtail possums from two regions in Australia (the first region is Victoria and the second is New South Wales and Queensland), where the possums may be considered a random sample from the population. We use logistic regression to classify the possums into the two regions. The outcome variable, called `pop`, takes value 1 when a possum is from Victoria and 0 when it is from New South Wales or Queensland.

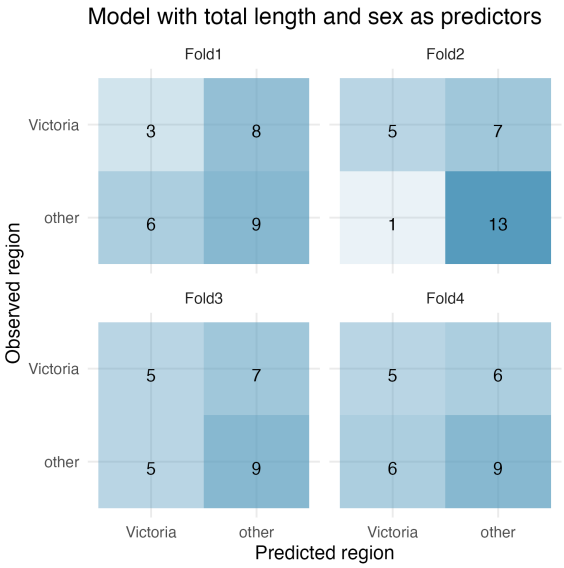
$$\log_e\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \times \texttt{tail\_l}$$

Model with tail length as predictor



$$\log_e\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \times \texttt{total\_l} + \beta_2 \times \texttt{sex}$$

Model with total length and sex as predictors



a. For the model with tail length, how many of the observations were correctly classified? What proportion of the observations were correctly classified?

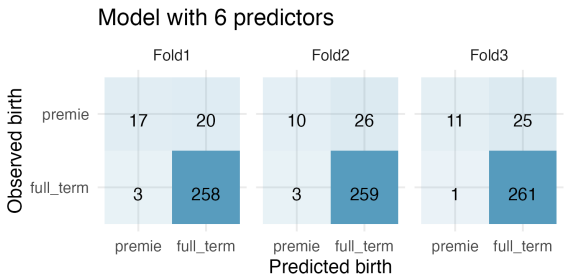
b. For the model with total length and sex, how many of the observations were correctly classified? What proportion of the observations were correctly classified?

c. If you have to choose between using only tail length as a predictor versus using total length and sex as predictors (for classification into region), which model would you choose? Explain.

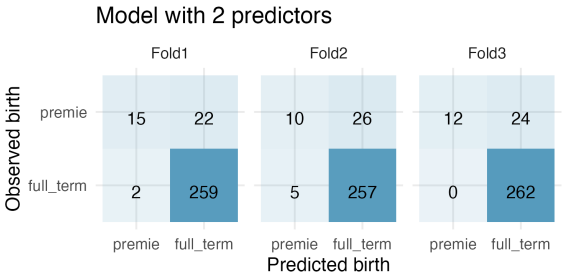
d. Given the predictions above, what third model might be preferable to either of the models above? Explain.

8. **Premature babies, cross-validation to choose model.** US Department of Health and Human Services, Centers for Disease Control and Prevention collect information on births recorded in the country. The data used here are a random sample of 1000 births from 2014 (with some rows removed due to missing data). We use logistic regression to model whether the baby is premature from various explanatory variables. (ICPSR 2014)

$$\log_e\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \times \texttt{mage} + \beta_2 \times \texttt{weight} + \beta_3 \times \texttt{mature} + \beta_4 \times \texttt{visits} + \beta_5 \times \texttt{gained} + \beta_6 \times \texttt{habit}$$



$$\log_e\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \times \texttt{weight} + \beta_2 \times \texttt{mature}$$



a. For the model with 6 predictors, how many of the observations were correctly classified? What proportion of the observations were correctly classified?

b. For the model with 2 predictors, how many of the observations were correctly classified? What proportion of the observations were correctly classified?

c. If you have to choose between the model with 6 predictors and the model with 2 predictors (for predicting whether a baby will be premature), which model would you choose? Explain.

Chapter 27

Applications: Model and infer

27.1 Case study: Mario Kart

In this case study, we consider Ebay auctions of a video game called *Mario Kart* for the Nintendo Wii. The outcome variable of interest is the total price of an auction, which is the highest bid plus the shipping cost. We will try to determine how total price is related to each characteristic in an auction while simultaneously controlling for other variables. For instance, all other characteristics held constant, are longer auctions associated with higher or lower prices? And, on average, how much more do buyers tend to pay for additional Wii wheels (plastic steering wheels that attach to the Wii controller) in auctions? Multiple regression will help us answer these and other questions.

The `mariokart` data can be found in the **openintro** R package.

The `mariokart` dataset includes results from 141 auctions. Four observations from this dataset are shown in Table 27.1, and descriptions for each variable are shown in Table 27.2. Notice that the condition and stock photo variables are **indicator variables**, similar to `bankruptcy` in the `loans` dataset from Chapter 25.

Table 27.1: Top four rows of the `mariokart` dataset.

price	cond_new	stock_photo	duration	wheels
51.5	new	yes	3	1
37.0	used	yes	7	1
45.5	new	no	3	1
44.0	new	yes	3	1

Table 27.2: Variables and their descriptions for the `mariokart` dataset.

Variable	Description
`price`	Final auction price plus shipping costs, in US dollars.
`cond_new`	Indicator variable for if the game is new (1) or used (0).
`stock_photo`	Indicator variable for if the auction's main photo is a stock photo.
`duration`	The length of the auction, in days, taking values from 1 to 10.
`wheels`	The number of Wii wheels included with the auction. A Wii wheel is an optional steering wheel accessory that holds the Wii controller.

27.1.1 Mathematical approach to linear models

In Table 27.3 we fit a mathematical linear regression model with the game's condition as a predictor of auction price.

$$E[\texttt{price}] = \beta_0 + \beta_1 \times \texttt{cond\_new}$$

Results of the model are summarized below:

Table 27.3: Summary of a linear model for predicting `price` based on `cond_new`.

term	estimate	std.error	statistic	p.value
(Intercept)	42.9	0.81	52.67	<0.0001
cond_new	10.9	1.26	8.66	<0.0001

GUIDED PRACTICE

Write down the equation for the model, note whether the slope is statistically different from zero, and interpret the coefficient.[1]

Sometimes there are underlying structures or relationships between predictor variables. For instance, new games sold on Ebay tend to come with more Wii wheels, which may have led to higher prices for those auctions. We would like to fit a model that includes all potentially important variables simultaneously, which would help us evaluate the relationship between a predictor variable and the outcome while controlling for the potential influence of other variables.

We want to construct a model that accounts for not only the game condition but simultaneously accounts for three other variables:

$$E[\texttt{price}] = \beta_0 + \beta_1 \times \texttt{cond\_new} + \beta_2 \times \texttt{stock\_photo} + \beta_3 \times \texttt{duration} + \beta_4 \times \texttt{wheels}$$

Table 27.4 summarizes the full model. Using the output, we identify the point estimates of each coefficient and the corresponding impact (measured with information on the standard error used to compute the p-value).

Table 27.4: Summary of a linear model for predicting `price` based on `cond_new`, `stock_photo`, `duration`, and `wheels`.

term	estimate	std.error	statistic	p.value
(Intercept)	36.21	1.51	23.92	<0.0001
cond_new	5.13	1.05	4.88	<0.0001
stock_photo	1.08	1.06	1.02	0.3085
duration	-0.03	0.19	-0.14	0.8882
wheels	7.29	0.55	13.13	<0.0001

GUIDED PRACTICE

Write out the model's equation using the point estimates from Table 27.4. How many predictors are there in the model? How many coefficients are estimated?[2]

[1] The equation for the line may be written as $\widehat{\texttt{price}} = 47.15 + 10.90 \times \texttt{cond\_new}$. Examining the regression output in Table 27.3 we can see that the p-value for `cond_new` is very close to zero, indicating there is strong evidence that the coefficient is different from zero when using this one-variable model. The variable `cond_new` is a two-level categorical variable that takes value 1 when the game is new and value 0 when the game is used. This means the 10.90 model coefficient predicts a price of an extra $10.90 for those games that are new versus those that are used.

GUIDED PRACTICE

What does β_4, the coefficient of variable x_4 (Wii wheels), represent? What is the point estimate of β_4?[3]

GUIDED PRACTICE

Compute the residual of the first observation in Table 27.1 using the equation identified in Table 27.4.[4]

EXAMPLE

In Table 27.3, we estimated a coefficient for `cond_new` in of $b_1 = 10.90$ with a standard error of $SE_{b_1} = 1.26$ when using simple linear regression. Why might there be a difference between that estimate and the one in the multiple regression setting?

If we examined the data carefully, we would see that there is **multicollinearity** among some predictors. For instance, when we estimated the connection of the outcome `price` and predictor `cond_new` using simple linear regression, we were unable to control for other variables like the number of Wii wheels included in the auction. That model was biased by the confounding variable `wheels`. When we use both variables, this particular underlying and unintentional bias is reduced or eliminated (though bias from other confounding variables may still remain).

27.1.2 Computational approach to linear models

Previously, using a mathematical model, we investigated the coefficients associated with `cond_new` when predicting `price` in a linear model.

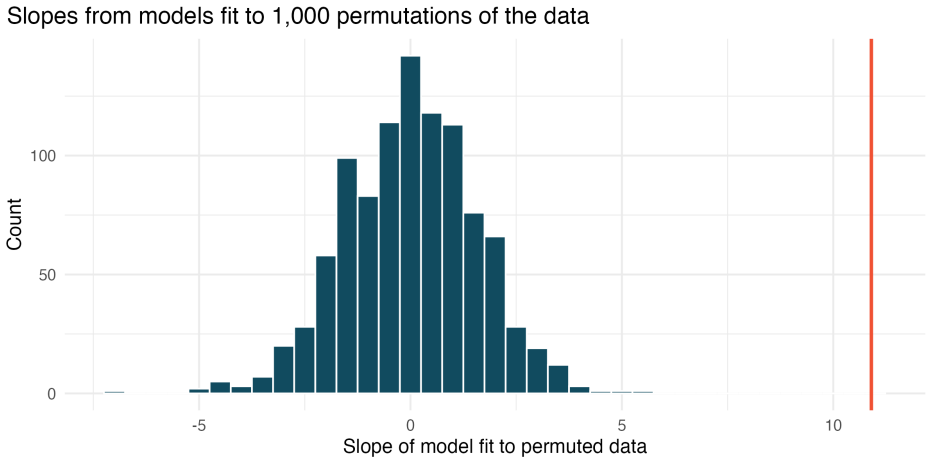


Figure 27.1: Estimated slopes from linear models (`price` regressed on `cond_new`) built on 1,000 randomized datasets. Each dataset was permuted under the null hypothesis.

[2] $\widehat{\texttt{price}} = 36.21 + 5.13 \times \texttt{cond\_new} + 1.08 \times \texttt{stock\_photo} - 0.03 \times \texttt{duration} + 7.29 \times \texttt{wheels}$, with 4 predictors but 5 coefficients (including the intercept).

[3] In the population of all auctions, it is the average difference in auction price for each additional Wii wheel included when holding the other variables constant. The point estimate is $b_4 = 7.29$

[4] $e_i = y_i - \hat{y}_i = 51.55 - 49.62 = 1.93$.

EXAMPLE

In Figure 27.1, the red line (the observed slope) is far from the bulk of the histogram. Explain why the randomly permuted datasets produce slopes that are quite different from the observed slope.

The null hypothesis is that, in the population, there is no linear relationship between the `price` and the `cond_new` of the *Mario Kart* games. When the data are randomly permuted, prices are randomly assigned to a condition (new or used), so that the null hypothesis is forced to be true, i.e., permutation is done under the assumption that no relationship between the two variables exists. In the actual study, the new *Mario Kart* games do actually cost more (on average) than the used games! So the slope describing the actual observed relationship is not one that is likely to have happened in a randomly dataset permuted under the assumption that the null hypothesis is true.

GUIDED PRACTICE

Using the histogram in Figure 27.1, find the p-value and conclude the hypothesis test in the context of the problem.[5]

GUIDED PRACTICE

Is the conclusion based on the histogram of randomized slopes consistent with the conclusion obtained using the mathematical model? Explain.[6]

Although knowing there is a relationship between the condition of the game and its price, we might be more interested in the difference in price, here given by the slope of the linear regression line. That is, β_1 represents the population value for the difference in price between new *Mario Kart* games and used games.

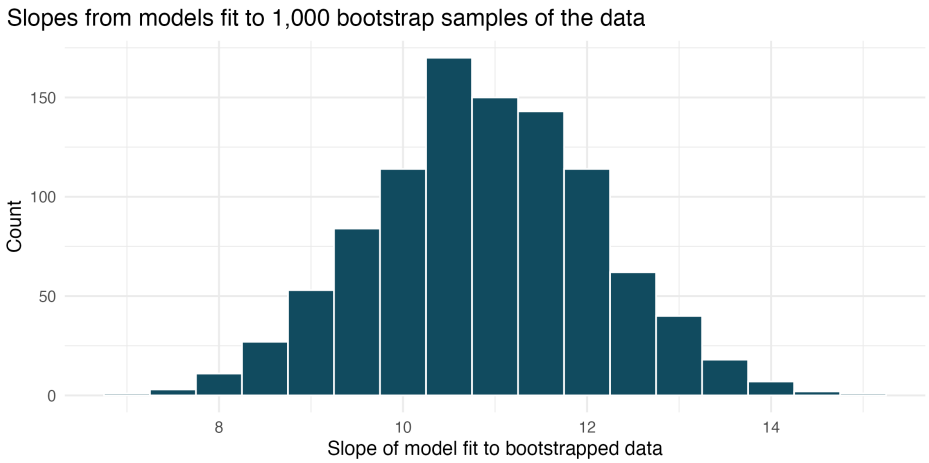


Figure 27.2: Estimated slopes from linear models (`price` regressed on `cond_new`) built on 1,000 bootstrapped datasets. Each bootstrap sample was taken from the original Mario Kart auction data.

[5]The observed slope is 10.9 which is nowhere near the range of values for the permuted slopes (roughly -5 to +5). Because the observed slope is not a plausible value under the null distribution, the p-value is essentially zero. We reject the null hypothesis and claim that there is a relationship between whether the game is new (or not) and the average predicted price of the game.

[6]The p-value in Table 27.3 is also essentially zero, so the null hypothesis is also rejected when the mathematical model approach is taken. Often, the mathematical and computational approaches to inference will give quite similar answers.

EXAMPLE

Figure 27.2 displays the slope estimates taken from bootstrap samples of the original data. Using the histogram, estimate the standard error of the slope. Is your estimate similar to the value of the standard error of the slope provided in the output of the mathematical linear model?

The slopes seem to vary from approximately 8 to 14. Using the empirical rule, we know that if a variable has a bell-shaped distribution, most of the observations will be with 2 standard errors of the center. Therefore, a rough approximation of the standard error is 1.5. The standard error given in Table 27.3 is 1.26 which is not too different from the value computed using the bootstrap approach.

GUIDED PRACTICE

Use Figure 27.2 to create a 90% standard error bootstrap confidence interval for the true slope. Interpret the interval in context.[7]

GUIDED PRACTICE

Use Figure 27.2 to create a 90% bootstrap percentile confidence interval for the true slope. Interpret the interval in context.[8]

27.1.3 Cross-validation

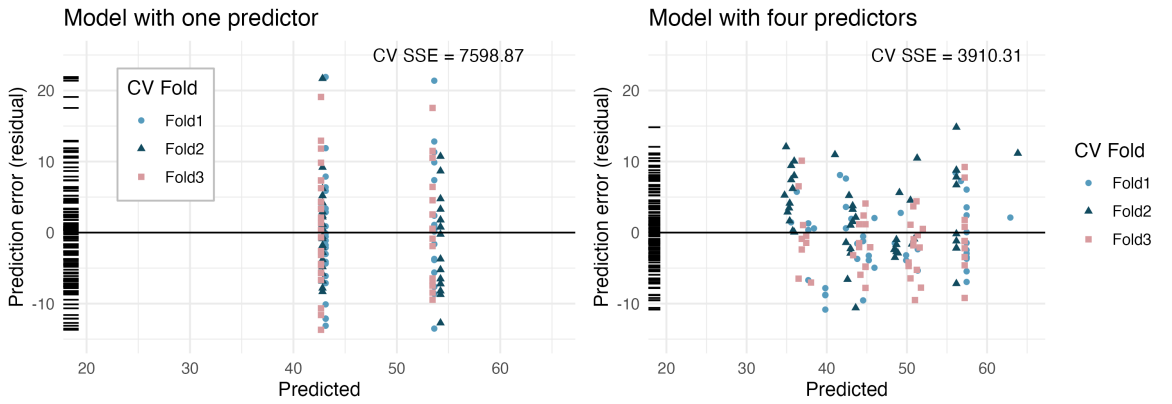
In Chapter 8, models were compared using R^2_{adj}. In Chapter 25, however, a computational approach was introduced to compare models by removing chunks of data one at a time and assessing how well the variables predicted the observations that had been held out.

Figure 27.3 was created by cross-validating models with the same variables as in Table 27.3 and Table 27.4. We applied 3-fold cross-validation, so 1/3 of the data was removed while 2/3 of the observations were used to build each model (first on `cond_new` only and then on `cond_new`, `stock_photo`, `duration`, and `wheels`). Note that each time 1/3 of the data is removed, the resulting model will produce slightly different model coefficients.

The points in Figure 27.3 represent the prediction (x-axis) and residual (y-axis) for each observation run through the cross-validated model. In other words, the model is built (using the other 2/3) without the observation (which is in the 1/3) being used. The residuals give us a sense for how well the model will do at predicting observations which were not a part of the original dataset, e.g., future studies.

[7] Using the bootstrap SE method, we know the normal percentile is $z^\star = 1.645$, which gives a CI of $b_1 \pm 1.645 \cdot SE \rightarrow 10.9 \pm 1.645 \cdot 1.5 \rightarrow (8.43, 13.37)$. For games that are new, the average price is higher by between \$8.43 and \$13.37 than games that are used, with 90% confidence.

[8] Because there were 1,000 bootstrap resamples, we look for the cutoffs which provide 50 bootstrap slopes on the left, 900 in the middle, and 50 on the right. Looking at the bootstrap histogram, the rough 95% confidence interval is \$9 to \$13.10. For games that are new, the average price is higher by between \$9.00 and \$13.10 than games that are used, with 90% confidence.



(a) `price` vs. `cond_new` **(b)** `price` vs. `cond_new`, `stock_photo`, `duration`, and `wheels`

Figure 27.3: Cross-validation predictions and errors from linear models built on two different sets of variables.

GUIDED PRACTICE

In Figure 27.3b, note the point at roughly predicted = 50 and prediction error = 10. Estimate the observed and predicted value for that observation.[9]

GUIDED PRACTICE

In Figure 27.3b, for the same point at roughly predicted = 50 and prediction error = 10, describe which cross-validation fold(s) were used to build its prediction model.[10]

GUIDED PRACTICE

By noting the spread of the cross-validated prediction errors (on the y-axis) in Figure 27.3, which model should be chosen for a final report on these data?[11]

GUIDED PRACTICE

Using the summary statistic cross-validation sum of squared errors (CV SSE), which model should be chosen for a final report on these data?[12]

[9] The predicted value is roughly $\widehat{\texttt{price}} = \$50$. The observed value is roughly $\texttt{price}_i = \$60$ riders (using $e_i = y_i - \hat{y}_i$).

[10] The point appears to be in fold 2, so folds 1 and 3 were used to build the prediction model.

[11] The cross-validated residuals on `cond_new` vary roughly from -15 to 15, while the cross-validated residuals on the four predictor model vary less, roughly from -10 to 10. Given the smaller residuals from the four predictor model, it seems as though the larger model is better.

[12] The CV SSE is smaller (by a factor of almost two!) for the model with four predictors. Using a single valued criterion (CV SSE) allows us to make a decision to choose the model with four predictors.

27.2 Interactive R tutorials

Navigate the concepts you've learned in this part in R using the following self-paced tutorials. All you need is your browser to get started!

Tutorial 6: Inferential modeling

https://openintrostat.github.io/ims-tutorials/06-model-infer

Tutorial 6 - Lesson 1: Inference in regression

https://openintro.shinyapps.io/ims-06-model-infer-01

Tutorial 6 - Lesson 2: Randomization test for slope

https://openintro.shinyapps.io/ims-06-model-infer-02

Tutorial 6 - Lesson 3: t-test for slope

https://openintro.shinyapps.io/ims-06-model-infer-03

Tutorial 6 - Lesson 4: Checking technical conditions for slope inference

https://openintro.shinyapps.io/ims-06-model-infer-04

Tutorial 6 - Lesson 5: Inference beyond the simple linear regression model

https://openintro.shinyapps.io/ims-06-model-infer-05

You can also access the full list of tutorials supporting this book at https://openintrostat.github.io/ims-tutorials.

27.3 R labs

Further apply the concepts you've learned in this part in R with computational labs that walk you through a data analysis case study.

Multiple linear regression - Grading the professor

https://www.openintro.org/go?id=ims-r-lab-model-infer

You can also access the full list of labs supporting this book at https://www.openintro.org/go?id=ims-r-labs.

Chapter A

Exercise solutions

A.1 Chapter 1

1. 23 observations and 7 variables.
3. (a) "Is there an association between air pollution exposure and preterm births?" (b) 143,196 births in Southern California between 1989 and 1993. (c) Measurements of carbon monoxide, nitrogen dioxide, ozone, and particulate matter less than $10\mu g/m^3$ (PM_{10}) collected at air-quality-monitoring stations as well as length of gestation. Continuous numerical variables.
5. (a) "What is the effect of gamification on learning outcomes compared to traditional teaching methods?" (b) 365 college students taking a statistics course (c) Gender (categorical), level of studies (categorical, ordinal), academic major (categorical), expertise in English language (categorical, ordinal), use of personal computers and games (categorical, ordinal), treatment group (categorical), score (numerical, discrete).
7. (a) Treatment: $10/43 = 0.23 \rightarrow 23\%$. (b) Control: $2/46 = 0.04 \rightarrow 4\%$. (c) A higher percentage of patients in the treatment group were pain free 24 hours after receiving acupuncture. (d) It is possible that the observed difference between the two group percentages is due to chance. (e) Explanatory: acupuncture or not. Response: if the patient was pain free or not.
9. (a) Experiment; researchers are evaluating the effect of fines on parents' behavior related to picking up their children late from daycare. (b) 10 cases: the daycare centers. (c) Number of late pickups (discrete numerical). (d) Week (numerical, discrete), group (categorical, nominal), number of late pickups (numerial, discrete), and study period (categorical, ordinal).
11. (a) 344 cases (penguins) are included in the data. (b) There are 4 numerical variables in the data: bill length, bill depth, and flipper length (measured in millimeters) and body mass (measured in grams). They are all continuous. (c) There are 3 categorical variables in the data: species (Adelie, Chinstrap, Gentoo), island (Torgersen, Biscoe, and Dream), and sex (female and male).
13. (a) Airport ownership status (public/private), airport usage status (public/private), region (Central, Eastern, Great Lakes, New England, Northwest Mountain, Southern, Southwest, Western Pacific), latitude, and longitude. (b) Airport ownership status: categorical, not ordinal. Airport usage status: categorical, not ordinal. Region: categorical, not ordinal. Latitude: numerical, continuous. Longitude: numerical, continuous.
15. (a) Year, number of baby girls named Fiona born in that year, nation. (b) Year (numerical, discrete), number of baby girls named Fiona born in that year (numerical, discrete), nation (categorical, nominal).
17. (a) County, state, driver's race, whether the car was searched or not, and whether the driver was arrested or not. (b) All categorical, non-ordinal. (c) Response: whether the car was searched or not. Explanatory: race of the driver.
19. (a) Observational study. (b) Dog: Lucy. Cat: Luna. (c) Oliver and Lily. (d) Positive, as the popularity of a name for dogs increases, so does the popularity of that name for cats.

A.2 Chapter 2

1. (a) Population mean, $\mu_{2007} = 52$; sample mean, $\bar{x}_{2008} = 58$. (b) Population mean, $\mu_{2001} = 3.37$; sample mean, $\bar{x}_{2012} = 3.59$.
3. (a) Population: all births, sample: 143,196 births between 1989 and 1993 in Southern California. (b) If births in this time span at the geography can be considered to be representative of all births, then the results are generalizable to the population of Southern California. However, since the study is observational the findings cannot be used to establish causal relationships.
5. (a) The population of interest is all college students studying statistics. The sample consists of 365 such students. (b) If the students in this sample, who are likely not randomly sampled, can be considered to be representative of all college students studying statistics, then the results are generalizable to the population defined above. This is probably not a reasonable assumption since these students are from two specific majors only. Additionally, since the study is experimental, the findings can be used to establish causal relationships.
7. (a) Observation. (b) Variable. (c) Sample statistic (mean). (d) Population parameter (mean).
9. (a) Observational. (b) Use stratified sampling to randomly sample a fixed number of students, say 10, from each section for a total sample size of 40 students.
11. (a) Positive, non-linear, somewhat strong. Countries in which a higher percentage of the population have access to the internet also tend to have higher average life expectancies, however rise in life expectancy trails off before around 80 years old. (b) Observational. (c) Wealth: countries with individuals who can widely afford the internet can probably also afford basic medical care. (Note: Answers may vary.)
13. (a) Simple random sampling is okay. In fact, it's rare for simple random sampling to not be a reasonable sampling method! (b) The student opinions may vary by field of study, so the stratifying by this variable makes sense and would be reasonable. (c) Students of similar ages are probably going to have more similar opinions, and we want clusters to be diverse with respect to the outcome of interest, so this would **not** be a good approach. (Additional thought: the clusters in this case may also have very different numbers of people, which can also create unexpected sample sizes.)
15. (a) The cases are 200 randomly sampled men and women. (b) The response variable is attitude towards a fictional microwave oven. (c) The explanatory variable is dispositional attitude. (d) Yes, the cases are sampled randomly, recruited online using Amazon's Mechanical Turk. (e) This is an observational study since there is no random assignment to treatments. (f) No, we cannot establish a causal link between the explanatory and response variables since the study is observational. (g) Yes, the results of the study can be generalized to the population at large since the sample is random.
17. (a) Simple random sample. Non-response bias, if only those people who have strong opinions about the survey responds their sample may not be representative of the population. (b) Convenience sample. Under coverage bias, their sample may not be representative of the population since it consists only of their friends. It is also possible that the study will have non-response bias if some choose to not bring back the survey. (c) Convenience sample. This will have a similar issues to handing out surveys to friends. (d) Multi-stage sampling. If the classes are similar to each other with respect to student composition this approach should not introduce bias, other than potential non-response bias.
19. (a) Exam performance. (b) Light level: fluorescent overhead lighting, yellow overhead lighting, no overhead lighting (only desk lamps). (c) Wearing glasses or not.
21. (a) Experiment. (b) Light level (overhead lighting, yellow overhead lighting, no overhead lighting) and noise level (no noise, construction noise, and human chatter noise). (c) Since the researchers want to ensure equal representation of those wearing glasses and not wearing glasses, wearing glasses is a blocking variable.
23. Need randomization and blinding. One possible outline: (1) Prepare two cups for each participant, one containing regular Coke and the other containing Diet Coke. Make sure the cups ar identical and contain equal amounts of soda. Label the cups (regular) and B (diet). (Be sure to randomize A and B for each trial!) (2) Give each participant the two cups, one cup at a time, in random order, and ask the participant to record a value that indicates ho much she liked the beverage. Be sure that neither the participant nor the person handing out the cups knows the identity of th beverage to make this a double-blind experiment. (Answers may vary.)
25. (a) Experiment. (b) Treatment: 25 grams of chia seeds twice a day, control: placebo. (c) Yes,

gender. (d) Yes, single blind since the patients were blinded to the treatment they received. (e) Since this is an experiment, we can make a causal statement. However, since the sample is not random, the causal statement cannot be generalized to the population at large.

27. (a) Non-responders may have a different response to this question, e.g., parents who returned the surveys likely don't have difficulty spending time with their children. (b) It is unlikely that the women who were reached at the same address 3 years later are a random sample. These missing responders are probably renters (as opposed to homeowners) which means that they might have a lower socio-economic status than the respondents. (c) There is no control group in this study, this is an observational study, and there may be confounding variables, e.g., these people may go running because they are generally healthier and/or do other exercises.
29. (a) Randomized controlled experiment. (b) Explanatory: treatment group (categorical, with 3 levels). Response variable: Psychological well-being. (c) No, because the participants were volunteers. (d) Yes, because it was an experiment. (e) The statement should say "evidence" instead of "proof".

A.3 Chapter 3

Application chapter, no exercises.

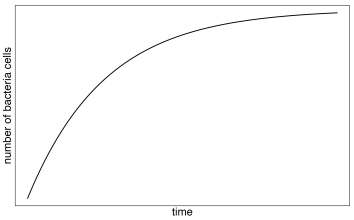
A.4 Chapter 4

1. (a) We see the order of the categories and the relative frequencies in the bar plot. (b) There are no features that are apparent in the pie chart but not in the bar plot. (c) We usually prefer to use a bar plot as we can also see the relative frequencies of the categories in this graph.
3. (a) The horizontal locations at which the age groups break into the various opinion levels differ, which indicates that likelihood of supporting protests varies by age group. Two variables may be associated. (b) Answers may vary. Political ideology/leaning and education level.
5. (a) Number of participants in each group. (b) Proportion of survival. (c) The standardized bar plot should be displayed as a way to visualize the survival improvement in the treatment versus the control group.
7. (a) The ridge plots do not tell us about the relationship between meat consumption and life expectancy. While it is true that the high income group of countries has highest meat consumption and highest life expectancy, we can't, for example, differentiate meat consumption across the low and middle income groups (so as to connect to life expectancy). Additionally, we don't know anything about the relationship betwen meat consumption and life expectancy *within* an income group. (b) When a relationship is confounded we cannot determine the causal mechanism. We don't know if the longer life expecancy is due to meat consumption or due to higher income (which comes with many other life-extending practices). (c) In order to investigate a specific confounding variable, first break the data into categories according to that confounding variable (here, income). Then look at the relationship of interest (here meat consumption and life expectancy) separately for each of the levels of the confounding variable (income).
9. (a) 41% of the JetBlue flights are delayed. 40.7% of the United Airlines flights are delayed. (b) For SFO: JetBlue had 39.7% delayed, United had 40% delayed (United had more delayed flights). For LAX: JetBlue had 40.1% delayed, United had 41% delayed (United had more delayed flights). For BQN: JetBlue had 45.7% delayed, United had 48.8% delayed (United had more delayed flights). (c) Note that JetBlue had substantially more flights than United out of BQN (where there was a high delay percentage). United had substantially more flights than United out of SFO and LAX, both of which had low delay percentages. So JetBlue's overall percentage delay is bumped up due to the BQN flights, and United's overall percentage delay is bumped down due to the SFO and LAX flights.

A.5 Chapter 5

1. (a) Positive association: mammals with longer gestation periods tend to live longer as well. (b) Association would still be positive. (c) No, they are not independent. See part (a).

3. The graph below shows a ramp up period. There may also be a period of exponential growth at the start before the size of the petri dish becomes a factor in slowing growth.



5. (a) Decrease: the new score is smaller than the mean of the 24 previous scores. (b) Calculate a weighted mean. Use a weight of 24 for the old mean and 1 for the new mean: $(24 \times 74 + 1 \times 64)/(24 + 1) = 73.6$. (c) The new score is more than 1 standard deviation away from the previous mean, so increase.

7. Any 10 employees whose average number of days off is between the minimum and the mean number of days off for the entire workforce at this plant.

9. (a) Dist B has a higher mean since $20 > 13$, and a higher standard deviation since 20 is further from the rest of the data than 13. (b) Dist A has a higher mean since $-20 > -40$, and Dist B has a higher standard deviation since -40 is farther away from the rest of the data than -20. (c) Dist B has a higher mean since all values in this Dist Are higher than those in Dist A, but both distribution have the same standard deviation since they are equally variable around their respective means. (d) Both distributions have the same mean since they're both centered at 300, but Dist B has a higher standard deviation since the observations are farther from the mean than in Dist A.

11. (a) About 26. (b) Since the distribution is right skewed the mean is higher than the median. (c) Q1: between 15 and 20, Q3: between 35 and 40, IQR: about 20. (d) Values that are considered to be unusually low or high lie more than 1.5×IQR away from the quartiles. Upper fence: Q3 + 1.5 × IQR = 37.5 + 1.5 × 20 = 67.5; Lower fence: Q1 - 1.5 × IQR = 17.5 + 1.5 × 20 = −12.5; The lowest AQI recorded is not lower than 5 and the highest AQI recorded is not higher than 65, which are both within the fences. Therefore none of the days in this sample would be considered to have an unusually low or high AQI.

13. The histogram shows that the distribution is bimodal, which is not apparent in the box plot. The box plot makes it easy to identify more precise values of observations outside of the whiskers.

15. (a) Right skewed, there is a natural boundary at 0 and only a few people have many pets. Center: median, variability: IQR. (b) Right skewed, there is a natural boundary at 0 and only a few people live a very long distance from work. Center: median, variability: IQR. (c) Symmetric. Center: mean, variability: standard deviation. (d) Left skewed. Center: median, variability: IQR. (e) Left skewed. Center: median, variability: IQR.

17. No, we would expect this distribution to be right skewed. There are two reasons for this: there is a natural boundary at 0 (it is not possible to watch less than 0 hours of TV) and the standard deviation of the distribution is very large compared to the mean.

19. No, the outliers are likely the maximum and the minimum of the distribution so a statistic based on these values cannot be robust to outliers.

21. The 75th percentile is 82.5, so 5 students will get an A. Also, by definition 25% of students will be above the 75th percentile.

23. (a) If $\frac{\bar{x}}{median} = 1$, then $\bar{x} = median$. This is most likely to be the case for symmetric distributions. (b) If $\frac{\bar{x}}{median} < 1$, then $\bar{x} < median$. This is most likely to be the case for left skewed distributions, since the mean is affected (and pulled down) by the lower values more so than the median. (c) If $\frac{\bar{x}}{median} > 1$, then $\bar{x} > median$. This is most likely to be the case for right skewed distributions, since the mean is affected (and pulled up) by the higher values more so than the median.

25. (a) The distribution of percentage of population that is Hispanic is extremely right skewed with majority of counties with less than 10% Hispanic residents. However there are a few counties

that have more than 90% Hispanic population. It might be preferable to, in certain analyses, to use the log-transformed values since this distribution is much less skewed. (b) The map reveals that counties with higher proportions of Hispanic residents are clustered along the Southwest border, all of New Mexico, a large swath of Southwest Texas, the bottom two-thirds of California, and in Southern Florida. In the map all counties with more than 40% of Hispanic residents are indicated by the darker shading, so it is impossible to discern how high Hispanic percentages go. The histogram reveals that there are counties with over 90% Hispanic residents. The histogram is also useful for estimating measures of center and spread. (c) Both visualizations are useful, but if we could only examine one, we should examine the map since it explicitly ties geographic data to each county's percentage.

A.6 Chapter 6

Application chapter, no exercises.

A.7 Chapter 7

1. (a) The residual plot will show randomly distributed residuals around 0. The variance is also approximately constant. (b) The residuals will show a fan shape, with higher variability for smaller x. There will also be many points on the right above the line. There is trouble with the model being fit here.
3. (a) Strong relationship, but a straight line would not fit the data. (b) Strong relationship, and a linear fit would be reasonable. (c) Weak relationship, and trying a linear fit would be reasonable. (d) Moderate relationship, but a straight line would not fit the data. (e) Strong relationship, and a linear fit would be reasonable. (f) Weak relationship, and trying a linear fit would be reasonable.
5. (a) Exam 2 since there is less of a scatter in the plot of course grade versus exam 2. Notice that the relationship between Exam 1 and the course grade appears to be slightly nonlinear. (b) (Answers may vary.) If Exam 2 is cumulative it might be a better indicator of how a student is doing in the class.
7. (a) $r = -0.7 \rightarrow (4)$. (b) $r = 0.45 \rightarrow (3)$. (c) $r = 0.06 \rightarrow (1)$. (d) $r = 0.92 \rightarrow (2)$.
9. (a) There is a moderate, positive, and linear relationship between shoulder girth and height. (b) Changing the units, even if just for one of the variables, will not change the form, direction or strength of the relationship between the two variables.
11. (a) There is a somewhat weak, positive, possibly linear relationship between the distance traveled and travel time. There is clustering near the lower left corner that we should take special note of. (b) Changing the units will not change the form, direction or strength of the relationship between the two variables. If longer distances measured in miles are associated with longer travel time measured in minutes, longer distances measured in kilometers will be associated with longer travel time measured in hours. (c) Changing units doesn't affect correlation: $r = 0.636$.
13. we can write the amount of meat consumption as an exact linear function of the amount of carbohydrate consumption. (a) $carbs = meat - 3$. (b) $carbs = meat + 2$. (c) $carbs = 2 \times meat$. Since the slopes are positive and these are perfect linear relationships, the correlation will be exactly 1 in all three parts. An alternative way to gain insight into this solution is to create a mock dataset, e.g., 5 countries with meat consumption of 10, 20, 50, 75, and 100 kg per capita, find the related carbohydrate consumption for each mock country, then create a scatterplot.
15. Correlation: no units. Intercept: cal. Slope: cal/cm.
17. Over-estimate. Since the residual is calculated as $observed - predicted$, a negative residual means that the predicted value is higher than the observed value.
19. (a) There is a positive, moderate, linear association between number of calories and amount of carbohydrates. In addition, the amount of carbohydrates is more variable for menu items with higher calories, indicating non-constant variance. There also appear to be two clusters of data: a patch of about a dozen observations in the lower left and a larger patch on the right side. (b) Explanatory: number of calories. Response: amount of carbohydrates (in grams). (c) With a regression line, we can predict the amount of carbohydrates for a given number of calories. This may be useful if only calorie counts for the food items are posted but the amount of carbohydrates in each food item is not readily available. (d) Food menu items with higher

predicted protein are predicted with higher variability than those without, suggesting that the model is doing a better job predicting protein amount for food menu items with lower predicted proteins.

21. (a) First calculate the slope: $b_1 = R \times s_y/s_x = 0.636 \times 113/99 = 0.726$. Next, make use of the fact that the regression line passes through the point $(\bar{x}, \bar{y})$: $\bar{y} = b_0 + b_1 \times \bar{x}$. Plug in $\bar{x}$, $\bar{y}$, and b_1, and solve for b_0: 51. Solution: $\widehat{travel\ time} = 51 + 0.726 \times distance$. (b) b_1: For each additional mile in distance, the model predicts an additional 0.726 minutes in travel time. b_0: When the distance travelled is 0 miles, the travel time is expected to be 51 minutes. It does not make sense to have a travel distance of 0 miles in this context. Here, the y-intercept serves only to adjust the height of the line and is meaningless by itself. (c) $R^2 = 0.636^2 = 0.40$. About 40% of the variability in travel time is accounted for by the model, i.e., explained by the distance travelled. (d) $\widehat{travel\ time} = 51 + 0.726 \times distance = 51 + 0.726 \times 103 \approx 126$ minutes. (Note: we should be cautious in our predictions with this model since we have not yet evaluated whether it is a well-fit model.) (e) $e_i = y_i - \hat{y}_i = 168 - 126 = 42$ minutes. A positive residual means that the model underestimates the travel time. (f) No, this calculation would require extrapolation.
23. (a) $\widehat{\texttt{poverty}} = 4.60 + 2.05 \times \texttt{unemployment\_rate}$. (b) The model predicts a poverty rate of 4.60% for counties with 0% unemployment, on average. This is not a meaningful value as no counties have such low unexmployment, it just serves to adjust the height of the regression line. (c) For each additional percentage increase in unemployment rate, poverty rate is predicted to be higher, on average, by 2.05%. (d) Unemployment rate explains 46% of the variability in poverty levels in US counties. (e) $\sqrt{0.46} = 0.678$.
25. (a) There is an outlier in the bottom right. Since it is far from the center of the data, it is a point with high leverage. It is also an influential point since, without that observation, the regression line would have a very different slope. (b) There is an outlier in the bottom right. Since it is far from the center of the data, it is a point with high leverage. However, it does not appear to be affecting the line much, so it is not an influential point. (c) The observation is in the center of the data (in the x-axis direction), so this point does *not* have high leverage. This means the point won't have much effect on the slope of the line and so is not an influential point.
27. (a) There is a negative, moderate-to-strong, somewhat linear relationship between percent of families who own their home and the percent of the population living in urban areas in 2010. There is one outlier: a state where 100% of the population is urban. The variability in the percent of homeownership also increases as we move from left to right in the plot. (b) The outlier is located in the bottom right corner, horizontally far from the center of the other points, so it is a point with high leverage. It is an influential point since excluding this point from the analysis would greatly affect the slope of the regression line.
29. (a) True. (b) False, correlation is a measure of the linear association between any two numerical variables.
31. (a) $r = 0.7 \rightarrow$ (1) (b) $r = 0.09 \rightarrow$ (4) (c) $r = -0.91 \rightarrow$ (2) (d) $r = 0.96 \rightarrow$ (3).

A.8 Chapter 8

1. Annika is right. All variables being highly correlated, including the predictor variables being highly correlated with each other, is not desirable as this would result in multicollinearity.
3. (a) The association between meat consumption and life expectancy is positive, moderate, and curved. (b) While tempting to say that eating meat may lead to a longer life expectancy, we do not have any sense of why the variables are associated. We are better off thinking that the countries with high meat consumption and high life expectancy are similar in many other ways (e.g., income bracket). (c) Within an income bracket, the relationship between meat consumption and life expectancy is not nearly as strong (as compared to when the data are aggregated into one plot).
5. No, they shouldn't include all variables as `days_since_start` and `days_since_race` are perfectly correlated with each other. They should only include one of them.
7. (a) $\widehat{\texttt{weight}} = 7.270 - 0.593 \times \texttt{habit}_{\texttt{smoker}}$. (b) The estimated body weight of babies born to smoking mothers is 0.593 pounds lower than those who are born to non-smoking mothers. Smoker: $\widehat{\texttt{weight}} = 7.270 - 0.593 \times 1 = 6.68$ pounds. Non-smoker: $\widehat{\texttt{weight}} = 7.270 - 0.593 \times 0 = 7.270$ pounds.
9. (a) Horror movies. (b) Not necessarily, the change in adjusted R^2 is quite small.

11. (a) $\widehat{\texttt{weight}} = -3.82 + 0.26 \times \texttt{weeks} + 0.02 \times \texttt{mage} + 0.37 \times \texttt{sex}_{\texttt{male}} + 0.02 \times \texttt{visits} - 0.43 \times \texttt{habit}_{\texttt{smoker}}$. (b) $b_{\texttt{weeks}}$: The model predicts a 0.26 pound increase in the birth weight of the baby for each additional week in length of pregnancy, all else held constant. $b_{\texttt{habit}_{\texttt{smoker}}}$: The model predicts a 0.43 pound decrease in the birth weight of the babies born to smoker mothers compared to non-smokers, all else held constant. (c) Habit might be correlated with one of the other variables in the model, which introduces multicollinearity and complicates model estimation. (d) -0.17~lbs.
13. Remove `gained`.
15. Add `weeks`.

A.9 Chapter 9

1. (a) False. The line is fit to predict the probability of success, not the binary outcome. (b) False. Residuals are not used in logistic regression like they are in linear regression because the observed value is always either zero or one (and the predicted value is a probability). The goal of the logistic regression is not to get a perfect prediction (of zero or one), so minimizing residuals is not part of the modeling process. (c) True.
3. (a) There are a few potential outliers, e.g., on the left in the variable total length, but nothing that will be of serious concern in a dataset this large. (b) When coefficient estimates are sensitive to which variables are included in the model, this typically indicates that some variables are collinear. For example, a possum's gender may be related to its head length, which would explain why the coefficient for `sex` changed when we removed the variable. Likewise, a possum's skull width is likely to be related to its head length and probably even much more closely related than the head length was to gender.
5. (a) The logistic model relating $\hat{p}$ to the predictors may be written as $\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = 33.5095 - 1.4207 \times \texttt{sex}_{\texttt{male}} - 0.2787 \times \texttt{skull\_w} + 0.5687 \times \texttt{total\_l} - 1.8057 \times \texttt{tail\_l}$. Only `total_l` has a positive association with a possum being from Victoria. (b) $\hat{p} = 0.0062$. While the probability is very near zero, we have not run diagnostics on the model. We might also be a little skeptical that the model will remain accurate for a possum found in a US zoo. For example, perhaps the zoo selected a possum with specific characteristics but only looked in one region. On the other hand, it is encouraging that the possum was caught in the wild. (Answers regarding the reliability of the model probability will vary.)
7. (a) The variable `exclaim_subj` should be removed, since it's removal reduces AIC the most (and the resulting model has lower AIC than the None Dropped model). (b) The variable `cc` should be removed. (c) Removing any variable will increase AIC, so we should not remove any variables from this set.
9. (a) The AIC is smallest using the variables `sex`, `head_l`, `skull_w`, `total_l`, and `tail_l` to predict region (AIC = 83.52), so we would choose that model. (b) If the metric is equivalent across two models with different numbers of variables, we usually want the model with smaller number of variables. Sometimes refered to as Occam's razor, the simplest explanation is often the one that will generalize most effectively.

A.10 Chapter 10

Application chapter, no exercises.

A.11 Chapter 11

1. (a) Mean. Each student reports a numerical value: a number of hours. (b) Mean. Each student reports a number, which is a percentage, and we can average over these percentages. (c) Proportion. Each student reports Yes or No, so this is a categorical variable and we use a proportion. (d) Mean. Each student reports a number, which is a percentage like in part (b). (e) Proportion. Each student reports whether s/he expects to get a job, so this is a categorical variable and we use a proportion.
3. (a) Alternative. (b) Null. (c) Alternative. (d) Alternative. (e) Null. (f) Alternative. (g) Null.
5. (a) $H_0 : \mu = 8$ (On average, New Yorkers sleep 8 hours a night.) $H_A : \mu < 8$ (On average, New Yorkers sleep less than 8 hours a night.) (b) $H_0 : \mu = 15$ (The average amount of company

time each employee spends not working is 15 minutes for March Madness.) $H_A : \mu > 15$ (The average amount of company time each employee spends not working is greater than 15 minutes for March Madness.)

7. (a) (i) False. Instead of comparing counts, we should compare percentages of people in each group who suffered cardiovascular problems. (ii) True. (iii) False. Association does not imply causation. We cannot infer a causal relationship based on an observational study. The difference from part (ii) is subtle. (iv) True. (b) Proportion of all patients who had cardiovascular problems: $\frac{7,979}{227,571} \approx 0.035$ (c) The expected number of heart attacks in the Rosiglitazone group, if having cardiovascular problems and treatment were independent, can be calculated as the number of patients in that group multiplied by the overall cardiovascular problem rate in the study: $67,593 * \frac{7,979}{227,571} \approx 2370$. (d) (i) H_0: The treatment and cardiovascular problems are independent. They have no relationship, and the difference in incidence rates between the Rosiglitazone and Pioglitazone groups is due to chance. H_A: The treatment and cardiovascular problems are not independent. The difference in the incidence rates between the Rosiglitazone and Pioglitazone groups is not due to chance and Rosiglitazone is associated with an increased risk of serious cardiovascular problems. (ii) A higher number of patients with cardiovascular problems than expected under the assumption of independence would provide support for the alternative hypothesis as this would suggest that Rosiglitazone increases the risk of such problems. (iii) In the actual study, we observed 2,593 cardiovascular events in the Rosiglitazone group. In the 100 simulations under the independence model, the simulated differences were never so high, which suggests that the actual results did not come from the independence model. That is, the variables do not appear to be independent, and we reject the independence model in favor of the alternative. The study's results provide convincing evidence that Rosiglitazone is associated with an increased risk of cardiovascular problems.

A.12 Chapter 12

1. (a) The statistic is the sample proportion (0.289); the parameter is the population proportion (unknown). (b) $\hat{p}$ and p. (c) Bootstrap sample proportion. (d) 0.289. (e) Roughly (0.22, 0.35). (f) We can be 90% confident that between 22% and 35% of all YouTube videos take place outdoors.
3. With 98% confidence, the true proportion of all US adults (in 2022) who get news from social media sometimes or often is between 0.487 and 0.51.
5. (a) A or perhaps D. (b) A, B, C, or D. (c) B or C. (d) B. (e) None.
7. (a) This claim is reasonable, since the entire interval lies above 50%. (b) The value of 70% lies outside of the interval, so we have convincing evidence that the researcher's conjecture is wrong. (c) A 90% confidence interval will be narrower than a 95% confidence interval. Even without calculating the interval, we can tell that 70% would not fall in the interval, and we would reject the researcher's conjecture based on a 90% confidence level as well.

A.13 Chapter 13

1. (a) 0.089 (b) 0.069 (c) 0.589 (d) $P(|Z| > 2) = P(Z < -2) + P(Z > 2)$ 0.046
3. (a) Verbal: $N(\mu = 151, \sigma = 7)$, Quant: $N(\mu = 153, \sigma = 7.67)$. (b) $Z_{VR} = 1.29$, $Z_{QR} = 0.52$. (c) She scored 1.29 standard deviations above the mean on the Verbal Reasoning section and 0.52 standard deviations above the mean on the Quantitative Reasoning section.

(d) She did better on the Verbal Reasoning section since her Z score on that section was higher. (e)$Perc_{VR} = 0.9007 \approx 90\%$, $Perc_{QR} = 0.6990 \approx 70\%$. (f) $100\% - 90\% = 10\%$ did better than her on VR, and $100\% - 70\% = 30\%$ did better than her on QR. (g) We cannot compare the raw scores since they are on different scales. Comparing her percentile scores is more appropriate when comparing her performance to others. (h) Answer to part (b) would not change as Z scores can be calculated for distributions that are not normal. However, we could not answer parts (d)-(f) since we cannot use the normal probability table to calculate probabilities and percentiles without a normal model.

1. (a) $Z = 0.84$, which corresponds to approximately 159 on QR. (b) $Z = -0.52$, which corresponds to approximately 147 on VR.

3. (a) $Z = 1.2$, $P(Z > 1.2) = 0.1151$. (b) $Z = -1.28 \rightarrow 70.6$◦F or colder.
5. (a) $N(25, 2.78)$. (b) $Z = 1.08$, $P(Z > 1.08) = 0.1401$. (c)The answers are very close because only the units were changed. (The only reason why they differ at all because 28° C is 82.4° F, not precisely 83° F.) (d) Since $IQR = Q3 - Q1$, we first need to find $Q3$ and $Q1$ and take the difference between the two. Remember that $Q3$ is the 75^{th} and $Q1$ is the 25^{th} percentile of a distribution. Q1 = 23.13, Q3 = 26.86, IQR = 26. 86 - 23.13 = 3.73.
7. (a) Recall that the general formula is $point\ estimate \pm z^\star \times SE$. First, identify the three different values. The point estimate is 45%, $z^\star = 1.96$ for a 95% confidence level, and $SE = 1.2\%$. Then, plug the values into the formula: $45\% \pm 1.96 \times 1.2\% \quad \rightarrow \quad (42.6\%, 47.4\%)$ We are 95% confident that the proportion of US adults who live with one or more chronic conditions is between 42.6% and 47.4%. (b) (i) False. Confidence intervals provide a range of plausible values, and sometimes the truth is missed. A 95% confidence interval "misses" about 5% of the time. (ii) True. Notice that the description focuses on the true population value. (iii) True. If we examine the 95% confidence interval, we can see that 50% is not included in this interval. This means that in a hypothesis test, we would reject the null hypothesis that the proportion is 0.5. (iv) False. The standard error describes the uncertainty in the overall estimate from natural fluctuations due to randomness, not the uncertainty corresponding to individuals' responses.
9. A Z score of 0.47 denotes that the sample proportion is 0.47 standard errors greater than the hypothesized value of the population proportion.
11. (a) Sampling distribution. (b) To know whether the distribution is skewed, we need to know the proportion. We've been told the proportion is likely above 5% and below 30%, and the success-failure condition would be satisfied for any of these values. If the population proportion is in this range, the sampling distribution will be symmetric. (c) Standard error. (d) The distribution will tend to be more variable when we have fewer observations per sample.

A.14 Chapter 14

1. (a) H_0: Anti-depressants do not affect the symptoms of Fibromyalgia. H_A: Anti-depressants do affect the symptoms of Fibromyalgia (either helping or harming). (b) Concluding that anti-depressants either help or worsen Fibromyalgia symptoms when they actually do neither. (c) Concluding that anti-depressants do not affect Fibromyalgia symptoms when they actually do.
3. (a) Scenario (i) is higher. Recall that a sample mean based on less data tends to be less accurate and have larger standard errors. (b) Scenario (i) is higher. The higher the confidence level, the higher the corresponding margin of error. (c) They are equal. The sample size does not affect the calculation of the p-value for a given Z score. (d) Scenario (i) is higher. If the null hypothesis is harder to reject (lower α), then we are more likely to make a Type II error when the alternative hypothesis is true.
5. The hypotheses should be about the population proportion (p), not the sample proportion. The null hypothesis should have an equal sign. The alternative hypothesis should have a not-equals sign, and it should reference the null value, $p_0 = 0.6$, not the observed sample proportion. The correct way to set up these hypotheses is: $H_0 : p = 0.6$ and $H_A : p \neq 0.6$.
7. Regardless of whether the students were making 95% intervals or 90% intervals, seven students with intervals that miss π seems totally reasonable. The students should not be docked for the intervals missing π.
9. True. If the sample size gets ever larger, then the standard error will become ever smaller. Eventually, when the sample size is large enough and the standard error is tiny, we can find statistically discernible yet very small differences between the null value and point estimate (assuming they are not exactly equal).

A.15 Chapter 15

Application chapter, no exercises.

A.16 Chapter 16

1. First, the hypotheses should be about the population proportion (p), not the sample proportion. Second, the null value should be what we are testing (0.25), not the observed value (0.29). The correct way to set up these hypotheses is: $H_0 : p = 0.25$ and $H_A : p > 0.25$.
3. (a) $H_0 : p = 0.20$, $H_A : p > 0.20$. (b) $\hat{p} = 159/650 = 0.245$. (c) Answers will vary. Each student can be represented with a card. Take 100 cards, 20 black cards representing those who support proposals to defund police departments and 80 red cards representing those who do not. Shuffle the cards and draw with replacement (shuffling each time in between draws) 650 cards representing the 650 respondents to the poll. Calculate the proportion of black cards in this sample, $\hat{p}_{sim}$, i.e., the proportion of those who upport proposals to defund police departments. The p-value will be the proportion of simulations where $\hat{p}_{sim} \geq 0.245$. (Note: We would generally use a computer to perform the simulations.) (d) There 1 only one simulated proportion that is at least 0.245, therefore the approximate p-value is 0.001. Your p-value may vary slightly since it is based on a visual estimate. Since the p-value is smaller than 0.05, we reject H_0. The data provide convincing evidence that the proportion of Seattle adults who support proposals to defund police departments is greater than 0.20, i.e., more than one in five.
5. (a) $H_0 : p = 0.5$, $H_A : p \neq 0.5$. (b) The p-value is roughly 0.4, There is not evidence in the data (possibly because there are only 7 cats being measured!) to conclude that the cats have a preference one way or the other between the two shapes.
7. (a) $SE(\hat{p}) = 0.189$. (c) Roughly 0.188. (c) Yes. (d) No. (e) The draws from the null hypothesis are discrete (only a few distinct options) and the mathematical model is continuous (infinite options on a continuum).
9. (a) The null hypothesis simulation was done with $p = 0.7$, and the data bootstrap simulation was done with $p = 0.6$. (b) The null hypothesis simulation is centered at 0.7; the data bootstrap is centered at 0.6. (c) The standard error of the sample proportion is given to be roughly 0.1 for both histograms. (d) Both histograms are reasonably symmetric. Note that histograms which describe the variability of proportions become more skewed as the center of the distribution gets closer to 1 (or zero) because the boundary of 1.0 restricts the symmetry of the tail of the distribution. For this reason, the null hypothesis simulation histogram is slightly more skewed (left).
11. (a) The null hypothesis simulation distribution for testing. The data bootstrap distribution for confidence intervals. (b) $H_0 : p = 0.7$; $H_A : p \neq 0.7$. p-value > 0.05. There is no evidence that the proportion of full-time statistics majors who work is different from 70%. (c) We are 98% confident that the true proportion of all full-time student statistics majors who work at least 5 hours per week is between 35% and 80%. (d) Using $z^\star = 2.33$, the 98% confidence interval is 0.367 to 0.833.
13. (a) False. Doesn't satisfy success-failure condition. (b) True. The success-failure condition is not satisfied. In most samples we would expect $\hat{p}$ to be close to 0.08, the true population proportion. While $\hat{p}$ can be much above 0.08, it is bound below by 0, suggesting it would take on a right skewed shape. Plotting the sampling distribution would confirm this suspicion. (c) False. $SE_{\hat{p}} = 0.0243$, and $\hat{p} = 0.12$ is only $\frac{0.12-0.08}{0.0243} = 1.65$ SEs away from the mean, which would not be considered unusual. (d) True. $\hat{p} = 0.12$ is 2.32 standard errors away from the mean, which is often considered unusual. (e) False. Decreases the SE by a factor of $1/\sqrt{2}$.
15. (a) True. See the reasoning of 6.1(b). (b) True. We take the square root of the sample size in the SE formula. (c) True. The independence and success-failure conditions are satisfied. (d) True. The independence and success-failure conditions are satisfied.
17. (a) False. A confidence interval is constructed to estimate the population proportion, not the sample proportion. (b) True. 95% CI: 82% $\pm$ 2%. (c) True. By the definition of the confidence level. (d) True. Quadrupling the sample size decreases the SE and ME by a factor of $1/\sqrt{4}$. (e) True. The 95% CI is entirely above 50%.
19. With a random sample, independence is satisfied. The success-failure condition is also satisfied. $ME = z^\star \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 1.96\sqrt{\frac{0.56 \times 0.44}{600}} = 0.0397 \approx 4\%$.
21. (a) No. The sample only represents students who took the SAT, and this was also an online survey. (b) (0.5289, 0.5711). We are 90% confident that 53% to 57% of high school seniors who took the SAT are fairly certain that they will participate in a study abroad program in college. (c) 90% of such random samples would produce a 90% confidence interval that includes the true

proportion. (d) Yes. The interval lies entirely above 50%.

23. (a) We want to check for a majority (or minority), so we use the following hypotheses: $H_0 : p = 0.5$ and $H_A : p \neq 0.5$. We have a sample proportion of $\hat{p} = 0.55$ and a sample size of $n = 617$ independents. Since this is a random sample, independence is satisfied. The success-failure condition is also satisfied: 617×0.5 and $617 \times (1 - 0.5)$ are both at least 10 (we use the null proportion $p_0 = 0.5$ for this check in a one-proportion hypothesis test). Therefore, we can model $\hat{p}$ using a normal distribution with a standard error of $SE = \sqrt{\frac{p(1-p)}{n}} = 0.02$. (We use the null proportion $p_0 = 0.5$ to compute the standard error for a one-proportion hypothesis test.) Next, we compute the test statistic: $Z = \frac{0.55-0.5}{0.02} = 2.5$. This yields a one-tail area of 0.0062, and a p-value of $2 \times 0.0062 = 0.0124$. Because the p-value is smaller than 0.05, we reject the null hypothesis. We have strong evidence that the support is different from 0.5, and since the data provide a point estimate above 0.5, we have strong evidence to support this claim by the TV pundit. (b) No. Generally we expect a hypothesis test and a confidence interval to align, so we would expect the confidence interval to show a range of plausible values entirely above 0.5. However, if the confidence level is misaligned (e.g., a 99% confidence level and a $\alpha = 0.05$ discernibility level), then this is no longer generally true.

25. (a) $H_0 : p = 0.5$. $H_A : p > 0.5$. Independence (random sample, $< 10\%$ of population) is satisfied, as is the success-failure conditions (using $p_0 = 0.5$, we expect 40 successes and 40 failures). $Z = 2.91 \rightarrow$ p- value $= 0.0018$. Since the p-value < 0.05, we reject the null hypothesis. The data provide strong evidence that the rate of correctly identifying a soda for these people is discernibly better than just by random guessing. (b) If in fact people cannot tell the difference between diet and regular soda and they randomly guess, the probability of getting a random sample of 80 people where 53 or more identify a soda correctly would be 0.0018.

27. (a) The sample is from all computer chips manufactured at the factory during the week of production. We might be tempted to generalize the population to represent all weeks, but we should exercise caution here since the rate of defects may change over time. (b) The fraction of computer chips manufactured at the factory during the week of production that had defects. (c) Estimate the parameter using the data: $\hat{p} = \frac{27}{212} = 0.127$. (d) *Standard error* (or SE). (e) Compute the SE using $\hat{p} = 0.127$ in place of p: $SE \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.127(1-0.127)}{212}} = 0.023$. (f) The standard error is the standard deviation of $\hat{p}$. A value of 0.10 would be about one standard error away from the observed value, which would not represent a very uncommon deviation. (Usually beyond about 2 standard errors is a good rule of thumb.) The engineer should not be surprised. (g) Recomputed standard error using $p = 0.1$: $SE = \sqrt{\frac{0.1(1-0.1)}{212}} = 0.021$. This value isn't very different, which is typical when the standard error is computed using relatively similar proportions (and even sometimes when those proportions are quite different!).

29. (a) The visitors are from a simple random sample, so independence is satisfied. The success-failure condition is also satisfied, with both 64 and $752-64 = 688$ above 10. Therefore, we can use a normal distribution to model $\hat{p}$ and construct a confidence interval. (b) The sample proportion is $\hat{p} = \frac{64}{752} = 0.085$. The standard error is $SE = \sqrt{\frac{0.085(1-0.085)}{752}} = 0.010$. (c) For a 90% confidence interval, use $z^\star = 1.65$. The confidence interval is $0.085 \pm 1.65 \times 0.010 \rightarrow (0.0685, 0.1015)$. We are 90% confident that 6.85% to 10.15% of first-time site visitors will register using the new design.

A.17 Chapter 17

1. (a) The parameter is $p_{Asican-Indian} - p_{Chinese}$. The statistic is $\hat{p}_{Asian-Indian} - \hat{p}_{Chinese} = 223/4373 - 279/4736 = -0.008$ (b) Roughly 0.005. (c) $H_0 : p_{Asian-Indian} - p_{Chinese} = 0;$, $H_A : p_{Asian-Indian} - p_{Chinese} \neq 0$. The evidence is borderline but worth further study. There is not strong evidence that the true difference in proportion of current smokers is different across the two ethnic groups.

3. (a) Roughly 0.00625. (b) We are 95% confident that the true proportion of Filipino Americans who are current smokers is between 5.28 and 7.72 percentage points higher in the control vaccine group than the proportion of Chinese Americans who smoke. (c) We are 95% confident that the true proportion of Filipino Americans who are current smokers is between 5.2 and 7.7 percentage points higher in the control vaccine group than the proportion of Chinese Americans who smoke.

5. (a) While the standard errors of the difference in proportion across the two graphs are roughly the same (approximately 0.012), the centers are not. Computational method A is centered at 0.07 (the difference in the observed sample proportions) and Computational method B is centered at 0. (b) What is the difference between the proportions of Bachelor's and Associate's students who believe that the COVID-19 pandemic will negatively impact their ability to complete the degree? (c) Is the proportion of Bachelor's students who believe that their ability to complete the degree will be negatively impacted by the COVID-19 pandemic different than that of Associate's students?
7. (a) 26 Yes and 94 No in Nevaripine and 10 Yes and 110 No in Lopinavir group. (b) $H_0 : p_N = p_L$. There is no difference in virologic failure rates between the Nevaripine and Lopinavir groups. $H_A : p_N \neq p_L$. There is some difference in virologic failure rates between the Nevaripine and Lopinavir groups. (c) Random assignment was used, so the observations in each group are independent. If the patients in the study are representative of those in the general population (something impossible to check with the given information), then we can also confidently generalize the findings to the population. The success-failure condition, which we would check using the pooled proportion ($\hat{p}_{pool} = 36/240 = 0.15$), is satisfied. $Z = 2.89 \rightarrow$ p-value $= 0.0039$. Since the p-value is low, we reject H_0. There is strong evidence of a difference in virologic failure rates between the Nevaripine and Lopinavir groups. Treatment and virologic failure do not appear to be independent.
9. (a) Standard error: $SE = \sqrt{\frac{0.79(1-0.79)}{347} + \frac{0.55(1-0.55)}{617}} = 0.03$. Using $z^\star = 1.96$, we get: $0.79 - 0.55 \pm 1.96 \times 0.03 \rightarrow (0.181, 0.299)$. We are 95% confident that the proportion of Democrats who support the plan is 18.1% to 29.9% higher than the proportion of Independents who support the plan. (b) True.
11. (a) In effect, we're checking whether men are paid more than women (or vice-versa), and we'd expect these outcomes with either chance under the null hypothesis: $H_0 : p = 0.5$ and $H_A : p \neq 0.5$. We'll use p to represent the fraction of cases where men are paid more than women. (b) There isn't a good way to check independence here since the jobs are not a simple random sample. However, independence doesn't seem unreasonable, since the individuals in each job are different from each other. The success-failure condition is met since we check it using the null proportion: $p_0 n = (1 - p_0)n = 10.5$ is greater than 10. We can compute the sample proportion, SE, and test statistic: $\hat{p} = 19/21 = 0.905$ and $SE = \sqrt{\frac{0.5 \times (1-0.5)}{21}} = 0.109$ and $Z = \frac{0.905 - 0.5}{0.109} = 3.72$. The test statistic Z corresponds to an upper tail area of about 0.0001, so the p-value is 2 times this value: 0.0002. Because the p-value is smaller than 0.05, we reject the notion that all these gender pay disparities are due to chance. Because we observe that men are paid more in a higher proportion of cases and we have rejected H_0, we can conclude that men are being paid higher amounts in ways not explainable by chance alone. If you're curious for more info around this topic, including a discussion about adjusting for additional factors that affect pay, please see the following video by Healthcare Triage: youtu.be/aVhgKSULNQA.
13. Before we can calculate a confidence interval, we must first check that the conditions are met. There aren't at least 10 successes and 10 failures in each of the four groups (treatment/control and yawn/not yawn), $(\hat{p}_C - \hat{p}_T)$ is not expected to be approximately normal and therefore cannot calculate a confidence interval for the difference between the proportions of participants who yawned in the treatment and control groups using large sample techniques and a critical Z score.
15. (a) False. The confidence interval includes 0. (b) False. We are 95% confident that 16% fewer to 2% Americans who make less than \$40,000 per year are not at all personally affected by the government shutdown compared to those who make \$40,000 or more per year. (c) False. As the confidence level decreases the width of the confidence level decreases as well. (d) True.
17. (a) Type I. (b) Type II. (c) Type II.
19. No. The samples at the beginning and at the end of the semester are not independent since the survey is conducted on the same students.
21. (a) The proportion of the normal curve centered at -0.1 with a standard deviation of 0.15 that is less than -2 * standard error is 0.09. (b) The proportion of the normal curve centered at -0.4 with a standard deviation of 0.145 that is less than 2 * standard error is 0.78. (c) The proportion of the normal curve centered at -0.1 with a standard deviation of 0.0671 that is less than 2 * standard error is 0.31. (d) The proportion of the normal curve centered at -0.4 with a standard deviation of 0.0678 that is less than 2 * standard error is 1. (e) The larger the value of δ and the larger the sample size, the more likely that the future study will lead to sample proportions

which are able to reject the null hypothesis.

A.18 Chapter 18

1. (a) Two-way table is shown below. (b-i) $E_{row_1,col_1} = \frac{(row\ 1\ total)\times(col\ 1\ total)}{table\ total} = 35$. This is lower than the observed value. (b-ii) $E_{row_2,col_2} = \frac{(row\ 2\ total)\times(col\ 2\ total)}{table\ total} = 115$. This is lower than the observed value.

	Quit		
Treatment	Yes	No	Total
Patch + support group	40	110	150
Only patch	30	120	150
Total	70	230	300

3. (a) Sun = 0.343, Partial = 0.325, Shade = 0.331. (b) For each, the numbers are listed in the order sun, partial, and shade: Desert (40,9, 38,7, 39.4), Mountain (36.7, 34.8, 35.5), Valley (36.4, 34.5, 35.1). (c) Yes. (d) We can't evaluate the association without a formal test.

5. The original dataset will have a higher Chi-squared statistic than the randomized dataset.

7. (a) The two variables are independent. (b) The randomized Chi-squared values range from zero to approximately 15. (c) The null hypothesis is that the variables are independent; the alternative hypothesis is that the variables are associated. The p-value is extremely small. The habitat provides information about the likelihood of being in the different sunshine states.

9. (a) The two variables are independent. (b) The randomized Chi-squared values range from zero to approximately 25. (c) The null hypothesis is that the variables are independent; the alternative hypothesis is that the variables are associated. The p-value is around 0. There is convincing evidence to claim that site and sunlight preference are associated. (d) With larger sample sizes, the power (the probability of rejecting H_0 when H_A is true) is higher.

11. (a) False. The Chi-square distribution has one parameter called degrees of freedom. (b) True. (c) True. (d) False. As the degrees of freedom increases, the shape of the Chi-square distribution becomes more symmetric.

13. The hypotheses are H_0 : Sleep levels and profession are independent. H_A : Sleep levels and profession are associated. The observations are independent and the sample sizes are large enough to conduct a Chi-square test of independence. The Chi-square statistic is 1 with 2 degrees of freedom. The p-value is 0.6. Since the p-value is high (default to alpha = 0.05), we fail to reject H_0. The data do not provide convincing evidence of an association between sleep levels and profession.

15. (a) H_0: The age of Los Angeles residents is independent of shipping carrier preference variable. H_A: The age of Los Angeles residents is associated with the shipping carrier preference variable. (b) The conditions are not satisfied since some expected counts are below 5.

A.19 Chapter 19

1. (a) Average sleep of 20 in sample vs. all New Yorkers. (b) Average height of students in study vs all undergraduates.
3. (a) Use the sample mean to estimate the population mean: 171.1. Likewise, use the sample median to estimate the population median: 170.3. (b) Use the sample standard deviation (9.4) and sample IQR ($177.8 - 163.8 = 14$). (c) $Z_{180} = 0.95$ and $Z_{155} = -1.71$. Neither of these observations is more than two standard deviations away from the mean, so neither would be considered unusual. (d) No, sample point estimates only estimate the population parameter, and they vary from one sample to another. Therefore we cannot expect to get the same mean

and standard deviation with each random sample. (e) We use the standard error of the mean to measure the variability in means of random samples of same size taken from a population. The variability in the means of random samples is quantified by the standard error. Based on this sample, $SE_{\bar{x}} = \frac{9.4}{\sqrt{507}} = 0.417$.

5. (a) The kindergartners will have a smaller standard deviation of heights. We would expect their heights to be more similar to each other compared to a group of adults' heights. (b) The standard error of the mean will depend on the variability of individual heights. The standard error of the adult sample averages will be around $9.4/\sqrt{100} = 0.94$cm. The standard error of the kindergartner sample averages will be smaller.
7. (a) $df = 6 - 1 = 5$, $t^\star_5 = 2.02$. (b) $df = 21 - 1 = 20$, $t^\star_{20} = 2.53$. (c) $df = 28$, $t^\star_{28} = 2.05$. (d) $df = 11$, $t^\star_{11} = 3.11$.
9. (a) 0.085, do not reject H_0. (b) 0.003, reject H_0. (c) 0.438, do not reject H_0. (d) 0.042, reject H_0.
11. (a) Roughly 0.1 weeks. (b) Roughly (38.45 weeks, 38.85 weeks). (c) Roughly (38.49 weeks, 38.91 weeks).
13. (a) False (b) False. (c) True. (d) False.
15. The mean is the midpoint: $\bar{x} = 20$. Identify the margin of error: $ME = 1.015$, then use $t^\star_{35} = 2.03$ and $SE = s/\sqrt{n}$ in the formula for margin of error to identify $s = 3$.
17. (a) H_0: $\mu = 8$ (New Yorkers sleep 8 hrs per night on average.) H_A: $\mu \neq 8$ (New Yorkers sleep less or more than 8 hrs per night on average.) (b) Independence: The sample is random. The min/max suggest there are no concerning outliers. $T = -1.75$. $df = 25 - 1 = 24$. (c) p-value $= 0.093$. If in fact the true population mean of the amount New Yorkers sleep per night was 8 hours, the probability of getting a random sample of 25 New Yorkers where the average amount of sleep is 7.73 hours per night or less (or 8.27 hours or more) is 0.093. (d) Since p-value > 0.05, do not reject H_0. The data do not provide strong evidence that New Yorkers sleep more or less than 8 hours per night on average. (e) Yes, since we did not rejected H_0.
19. With a larger critical value, the confidence interval ends up being wider. This makes intuitive sense as when we have a small sample size and the population standard deviation is unknown, we should have a wider interval than if we knew the population standard deviation, or if we had a large enough sample size.
21. (a) We will conduct a 1-sample t-test. H_0: $\mu = 5$. H_A: $\mu \neq 5$. We'll use $\alpha = 0.05$. This is a random sample, so the observations are independent. To proceed, we assume the distribution of years of piano lessons is approximately normal. $SE = 2.2/\sqrt{20} = 0.4919$. The test statistic is $T = (4.6 - 5)/SE = -0.81$. $df = 20 - 1 = 19$. The one-tail area is about 0.21, so the p-value is about 0.42, which is bigger than $\alpha = 0.05$ and we do not reject H_0. That is, we do not have sufficiently strong evidence to reject the notion that the average is 5 years. (b) Using $SE = 0.4919$ and $t^\star_{df=19} = 2.093$, the confidence interval is (3.57, 5.63). We are 95% confident that the average number of years a child takes piano lessons in this city is 3.57 to 5.63 years. (c) They agree, since we did not reject the null hypothesis and the null value of 5 was in the t-interval.

A.20 Chapter 20

1. The hypotheses should use population means (μ) not sample means ($\bar{x}$), the null hypothesis should set the two population means equal to each other, the alternative hypothesis should be two-tailed and use a not equal to sign.
3. $H_0 : \mu_{0.99} = \mu_1$ and $H_A : \mu_{0.99} \neq \mu_1$. p-value < 0.05, reject H_0. The data provide convincing evidence that the difference in population averages of price per carat of 0.99 carats and 1 carat diamonds are different.
5. (a) We are 95% confident that the population average price per carat of 0.99 carat diamonds is \$2 to \$23 lower than the population average price per carat of 1 carat diamonds. (b) We are 95% confident that the population average price per carat of 0.99 carat diamonds is \$2.91 to \$21.10 lower than the population average price per carat of 1 carat diamonds.
7. The difference is not zero (statistically discernible), but there is no evidence that the difference is large (practically important), because the interval provides values as low as 1 lb.
9. $H_0 : \mu_{0.99} = \mu_1$ and $H_A : \mu_{0.99} \neq \mu_1$. Independence: Both samples are random and represent less than 10% of their respective populations. Also, we have no reason to think that the 0.99 carats

are not independent of the 1 carat diamonds since they are both sampled randomly. Normality: The distributions are not extremely skewed, hence we can assume that the distribution of the average differences will be nearly normal as well. $T_{22} = -2.7$, p-value = 0.0131. Since p-value less than 0.05, reject H_0. The data provide convincing evidence that the difference in population averages of price per carat of 0.99 carats and 1 carat diamonds are different.

11. We are 95% confident that the population average price per carat of 0.99 carat diamonds is \$2.96 to \$22.42 lower than the population average price per carat of 1 carat diamonds.

13. (a) $\mu_{\bar{x}_1} = 15$, $\sigma_{\bar{x}_1} = 20/\sqrt{50} = 2.8284$. (b) $\mu_{\bar{x}_2} = 20$, $\sigma_{\bar{x}_1} = 10/\sqrt{30} = 1.8257$. (c) $\mu_{\bar{x}_2-\bar{x}_1} = 20 - 15 = 5$, $\sigma_{\bar{x}_2-\bar{x}_1} = \sqrt{\left(20/\sqrt{50}\right)^2 + \left(10/\sqrt{30}\right)^2} = 3.3665$. (d) Think of $\bar{x}_1$ and $\bar{x}_2$ as being random variables, and we are considering the standard deviation of the difference of these two random variables, so we square each standard deviation, add them together, and then take the square root of the sum: $SD_{\bar{x}_2-\bar{x}_1} = \sqrt{SD^2_{\bar{x}_2} + SD^2_{\bar{x}_1}}$.

15. (a) Chicken fed linseed weighed an average of 218.75 grams while those fed horsebean weighed an average of 160.20 grams. Both distributions are relatively symmetric with no apparent outliers. There is more variability in the weights of chicken fed linseed. (b) $H_0 : \mu_{ls} = \mu_{hb}$. $H_A : \mu_{ls} \neq \mu_{hb}$. We leave the conditions to you to consider. $T = 3.02$, $df = min(11, 9) = 9 \rightarrow$ p-value = 0.014. Since p-value < 0.05, reject H_0. The data provide strong evidence that there is a discernible difference between the average weights of chickens that were fed linseed and horsebean. (c) Type I error, since we rejected H_0. (d) Yes, since p-value > 0.01, we would not have rejected H_0.

17. $H_0 : \mu_C = \mu_S$. $H_A : \mu_C \neq \mu_S$. $T = 3.27$, $df = 11 \rightarrow$ p-value = 0.007. Since p-value < 0.05, reject H_0. The data provide strong evidence that the average weight of chickens that were fed casein is different than the average weight of chickens that were fed soybean (with weights from casein being higher). Since this is a randomized experiment, the observed difference can be attributed to the diet.

19. $H_0 : \mu_T = \mu_C$. $H_A : \mu_T \neq \mu_C$. $T = 2.24$, $df = 21 \rightarrow$ p-value = 0.036. Since p-value < 0.05, reject H_0. The data provide strong evidence that the average food consumption by the patients in the treatment and control groups are different. Furthermore, the data indicate patients in the distracted eating (treatment) group consume more food than patients in the control group.

A.21 Chapter 21

1. Paired, data are recorded in the same cities at two different time points. The temperature in a city at one point is not independent of the temperature in the same city at another time point

3. (a) Since it's the same students at the beginning and the end of the semester, there is a pairing between the datasets, for a given student their beginning and end of semester grades are dependent. (b) Since the subjects were sampled randomly, each observation in the men's group does not have a special correspondence with exactly one observation in the other (women's) group. (c) Since it's the same subjects at the beginning and the end of the study, there is a pairing between the datasets, for a subject student their beginning and end of semester artery thickness are dependent. (d) Since it's the same subjects at the beginning and the end of the study, there is a pairing between the datasets, for a subject student their beginning and end of semester weights are dependent.

5. False. While it is true that paired analysis requires equal sample sizes, only having the equal sample sizes isn't, on its own, sufficient for doing a paired test. Paired tests require that there be a special correspondence between each pair of observations in the two groups.

7. (a) Let $diff = 2022 - 1950$. Then the hypotheses are $H_0 : \mu_{diff} = 0$ and $H_A : \mu_{diff} \neq 0$. (b) The observed average of difference is outside the randomized differences. (c) Since the p-value < 0.05, reject H_0. There is evidence of a difference between the average 90^{th} percentile high temperature in 2022 and the average 90^{th} percentile high temperature in 1950.

9. (a) Roughly (1.5°F, 3.5°F). (b) Roughly (1.5°F, 3.56°F). (c) We are 90% confident that the true average of the difference in 90^{th} percentile high temperature in 2022 vs 1950 is somewhere between 1.5°F and 3.5°F. We are 90% confident that the true average of the difference in 90^{th} percentile high temperature in 2022 vs 1950 is somewhere between 1.5°F and 3.56°F. (d) There is a discernible difference.

11. (a) For each observation in the 1950 dataset, there is exactly one specially corresponding observation in the 2022 dataset for the same geographic location. The data are paired. (b) $H_0 : \mu_{\text{diff}} = 0$

(There is no difference in the 90^{th} percentile high temperature in 1950 and 2022 for NOAA stations.) $H_A : \mu_{\text{diff}} \neq 0$ (There is a difference.) (c) Locations were not randomly sampled across the geographic region, so we need to be careful concluding independence. However, the question above describes the data as representative of the land area of the lower 48 states, so independence is reasonable. The sample size is 26 which is close to 30, so we're just looking for particularly extreme outliers: none are present (the observation off to the right in the histogram would be considered a outlier, but not a particularly extreme one). Therefore, the conditions are reasonably satisfied. (d) $SE = 2.95/\sqrt{26} = 0.579$. $T = \frac{2.53-0}{0.579} = 4.37$ with degrees of freedom $df = 26 - 1 = 25$, which leads to a one-tail area of 0.0000954 and a p-value of about 0.0002. (e) Since the p-value is less than 0.05, we reject H_0. The data provide strong evidence that NOAA stations observed a hotter 90^{th} percentile high temperature in 2022 than in 1950. (f) Type I error, since we may have incorrectly rejected H_0. This error would mean that NOAA stations did not actually observe an increase, but the sample we took just so happened to make it appear that this was the case. (g) No, since we rejected H_0, which had a null value of 0.

13. (a) $SE = 0.579$ and $t^\star_{25} = 1.71$. $2.53 \pm 1.71 \times 0.579 \rightarrow (1.54\,^{F,3.52}\,F)$. (b) We are 90% confident that the true average of the difference in 90^{th} percentile high temperature in 2022 vs 1950 is somewhere between 1.54°F and 3.52°F. (c) Yes, since the interval lies entirely above 0.
15. (a) Each student study under each condition, use the difference in individual student scores. (b) Each student study under one condition, use the difference in average across the two conditions.
17. (a)$H_0 : \mu_{diff} = 0$. $H_A : \mu_{diff} \neq 0$. $T = -2.71$. $df = 5$. p-value = 0.042. Since p-value $<$ 0.05, reject H_0. The data provide strong evidence that the average number of traffic accident related emergency room admissions are different between Friday the 6^{th} and Friday the 13^{th}. Furthermore, the data indicate that the direction of that difference is that accidents are lower on Friday the 6^{th} relative to Friday the 13^{th}. (b) (-6.49, -0.17). (c) This is an observational study, not an experiment, so we cannot so easily infer a causal intervention implied by this statement. It is true that there is a difference. However, for example, this does not mean that a responsible adult going out on Friday the 13^{th} has a higher chance of harm than on any other night.

A.22 Chapter 22

1. Alternative.
3. (a) Means across original data are more variable. (b) Standard deviation of egg lengths are about the same for both plots. (c) F statistic is bigger for the original data.
5. H_0: $\mu_1 = \mu_2 = \cdots = \mu_6$. H_A: The average weight varies across some (or all) groups. Independence: Chicks are randomly assigned to feed types (presumably kept separate from one another), therefore independence of observations is reasonable. Approx. normal: the distributions of weights within each feed type appear to be fairly symmetric. Constant variance: Based on the side-by-side box plots, the constant variance assumption appears to be reasonable. There are differences in the actual computed standard deviations, but these might be due to chance as these are quite small samples. $F_{5,65} = 15.36$ and the p-value is approximately 0. With such a small p-value, we reject H_0. The data provide convincing evidence that the average weight of chicks varies across some (or all) feed supplement groups.
7. (a) H_0: The population mean of MET for each group is equal to the others. H_A: At least one pair of means is different. (b) Independence: We don't have any information on how the data were collected, so we cannot assess independence. To proceed, we must assume the subjects in each group are independent. In practice, we would inquire for more details. Normality: The data are bound below by zero and the standard deviations are larger than the means, indicating very strong skew. However, since the sample sizes are extremely large, even extreme skew is acceptable. Constant variance: This condition is sufficiently met, as the standard deviations are reasonably consistent across groups. (c) Since p-value is very small, reject H_0. The data provide convincing evidence that the average MET differs between at least one pair of groups.
9. (a) H_0: Average GPA is the same for all majors. H_A: At least one pair of means are different. (b) Since p-value $>$ 0.05, fail to reject H_0. The data do not provide convincing evidence of a difference between the average GPAs across three groups of majors. (c) The total degrees of freedom is $195 + 2 = 197$, so the sample size is $197 + 1 = 198$.
11. (a) False. As the number of groups increases, so does the number of comparisons and hence the modified discernibility level decreases. (b) True. (c) True. (d) False. We need observations to

be independent regardless of sample size.

13. (a) Left is Dataset B. (b) Right is Dataset A.

A.23 Chapter 23

Application chapter, no exercises.

A.24 Chapter 24

1. (a) $H_0 : \beta_1 = 0$, $H_A : \beta_1 \neq 0$. (b) The observed slope of 0.604 is not a plausible value, the p-value is extremely small, and the null hypothesis can be rejected. c. The p-value is also extremely small.
3. (a) The relationship is positive, moderate-to-strong, and linear. There are a few outliers but no points that appear to be influential. (b) $\widehat{\texttt{wgt}} = -105.0113 + 1.0176 \times \texttt{hgt}$. Slope: For each additional centimeter in height, the model predicts the average weight to be 1.0176 additional kilograms (about 2.2 pounds). Intercept: People who are 0 centimeters tall are expected to weigh -105.0113 kilograms. This is obviously not possible. Here, the y- intercept serves only to adjust the height of the line and is meaningless by itself. (c) H_0: The true slope coefficient of height is zero ($\beta_1 = 0$). H_A: The true slope coefficient of height is different than zero ($\beta_1 \neq 0$). The p-value for the two-sided alternative hypothesis ($\beta_1 \neq 0$) is incredibly small, so we reject H_0. The data provide convincing evidence that height and weight are positively correlated. The true slope parameter is indeed greater than 0. (d) $R^2 = 0.72^2 = 0.52$. Approximately 52% of the variability in weight can be explained by the height of individuals.
5. (a) Roughly 0.53 to 0.67. (b) For individuals with one cm larger shoulder girth, their average height is predicted to be between 0.53 and 0.67 cm taller, with 98% confidence.
7. (a) Approximately 0.025. (b) $b_1 \pm 2.33 \times SE \rightarrow (0.546, 0.662)$. (c) For individuals with one cm larger shoulder girth, their average height is predicted to be between 0.546 and 0.662 cm taller, with 98% confidence.
9. (a) $r = \sqrt{0.518} \approx +0.72$. We know the correlation is positive due to the positive association between the variables seen in the scatterplot (above in previous exercise). (b) The residuals have a larger spread above the horizontal line at zero than below the line at zero. This indicates that the values are not symmetric around zero (so therefore not normally distributed). However, the violation is not extreme, and a simple least squares fit is probably appropriate for these data.
11. (a) $H_0 : \beta_1 = 0$, $H_A : \beta_1 \neq 0$. (b) The observed slope of 2.559 is not a plausible value, the p-value is extremely small, and the null hypothesis can be rejected. (c) The p-value is also extremely small.
13. (a) Rough 90% confidence interval is 1.9 to 3.1. (b) For a one unit (one percentage point) increase in poverty across given metropolitan areas, the predicted average annual murder rate will be between 1.9 and 3.1 persons per million larger, with 90% confidence.
15. (a) $r = \sqrt{0.706} \approx +0.84$. We know the correlation is positive due to the positive association shown in the scatterplot. (b) The technical conditions all seem to be met.
17. (a) With only sixteen observations in the analysis there are not enough data points to establish any patterns in the residual plot. That said, the sixteen observations do not show any large deviations of LNE conditions. We do not know if the volunteers were friends, for example, which would violate the independence condition. (b) The layout of the points does not indicate any deviation form the LINE technical conditions. The small number of points, however, suggests that care should be given to making sure that the individuals in the study are a good representative sample of the population to which we would like to infer the results.
19. (a) The **L**inearity and **N**ormality conditions seem to be met. If anything, the **E**qual variance condition is violated due to the a fan shaped pattern in the plot, which indicates non-constant variability in the residuals (little variability when x is small, more variability when x is large). We do not know if the cats were randomly sampled (i.e., are independent from one another), but we have no reason to believe that they are not independent. (b) Unequal variability does not affect the fit of the line. The line will continue to model the average heart weight of cats at a given body weight. However, the p-value for the inference on the line will be affected by the unequal variability. How much? Probably not much given that the violation is quite minimal.

A.25 Chapter 25

1. (a) (-0.044, 0.346). We are 95% confident that student who go out more than two nights a week on average have GPAs 0.044 points lower to 0.346 points higher than those who do not go out more than two nights a week, when controlling for the other variables in the model. (b) Yes, since the p-value is larger than 0.05 in all cases (not including the intercept).
3. (a) `volume` and `diam`; `volume` and `height`; `diam` and `height`. (b) Each is discernible in its own model. (c) When both `diam`eter and `height` are used in the multiple linear regression model, both continue to be discernible predictors of `volume`.
5. (a) **L**inearity: Horror movies seem to show a much different pattern than the other genres. While the residuals plots show a random scatter over years and in order of data collection, there is a clear pattern in residuals for various genres, which signals that this regression model is not appropriate for these data. **I**ndependent observations: The variability of the residuals is higher for data that comes later in the dataset. We don't know if the data are sorted by year, but if so, there may be a temporal pattern in the data that voilates the independence condition. **N**ormality: The residuals are right skewed (skewed to the high end). Constant or **E**qual variability: The residuals vs. predicted values plot reveals some outliers. This plot for only babies with predicted birth weights between 6 and 8.5 pounds looks a lot better, suggesting that for bulk of the data the constant variance condition is met.
7. (a) Linearity: With so many observations in the dataset, we look for particularly extreme outliers in the histogram of residuals and do not see any. We also don't see a non-linear pattern emerging in the residuals vs. predicted plot. Independent observations: The sample is random and there does not seem to be a trend in the residuals vs. order of data collection plot. Normality: The histogram of residuals appears to be unimodal and symmetic, centered at 0. Constant or equal variability: The residuals vs. predicted values plot reveals some outliers. This plot for only babies with predicted birth weights between 6 and 8.5 pounds looks a lot better, suggesting that for bulk of the data the constant variance condition is met. All concerns raised here are relatively mild. There are some outliers, but there is so much data that the influence of such observations will be minor. (b) H_0: The true slope coefficient of habit is zero ($\beta_5 = 0$). H_A: The true slope coefficient of height is different than zero ($\beta_5 \neq 0$). The p-value for the two-sided alternative hypothesis ($\beta_5 \neq 0$) is incredibly 0.0007 (smaller than 0.05), so we reject H_0. The data provide convincing evidence that height and weight are positively correlated, given the other variables in the model. The true slope parameter is indeed greater than 0.
9. (a) Roughly $\widehat{\texttt{weight}} = 11$ pounds and $\texttt{weight}_i = 7$ pounds. (b) Folds 1, 2, and 4 were used to build the prediction model. (c) The plot on the top estimates 8 parameters; the plot on the bottom estimates 3 parameters. (d) The residuals are not substantially different.
11. (a) The plots are difficult to differentiate. (b) The CV SSE is smaller for the model with only two predictors. (c) The model with more predictors seems to be over-fitting the data used to model build at the expense of not fitting (as well) the cross-validation hold out set for prediction.

A.26 Chapter 26

1. No, logistic regression is not appropriate because the response (or outcome) variable is not binary. Linear regression is likely to be more appropriate.
3. $H_0 : \beta_1 = 0$, the slope of the model predicting kids' marijuana use in college from their parents' marijuana use in college is 0. $H_A : \beta_1 \neq 0$, the slope of the model predicting kids' marijuana use in college from their parents' marijuana use in college is different than 0. The test statistic is $Z = 4.09$ and the associated p-value is less than 0.0001. With a small p-value we reject H_0. The data provide convincing evidence that the slope of the model predicting kids' marijuana use in college from their parents' marijuana use in college is different than 0, i.e., that parents' marijuana use in college is a discernible predictor of kids' marijuana use in college.
5. (a) 26 observations are in Fold2. 8 correctly and 2 incorrectly predicted to be from Victoria. (b) 78 observations are used to build the model. (c) 2 coefficients for tail length; 3 coefficients for total length and sex.
7. (a) 76, 73.1%. (b) 58, 55.8%. (c) The tail length model should be chosen for classification purposes. (d) A model using all three predictors might be superior to either of the smaller models.

Chapter B

References

Adolph, S. C. 1987. "Physiological and Behavioral Ecology of the Lizards *Sceloporus Occidentalis* and *Sceloporus Graciosus*." PhD thesis, University of Washington, Seattle, Washington.

———. 1990. "Influence of Behavioral Thermoregulation on Microhabitat Use by Two *Sceloporus* Lizards." *Ecology* 71: 315–27.

Allais, G., M. Romoli, S. Rolando, G. Airola, I. Castagnoli Gabellari, R. Allais, and C. Benedetto. 2011. "Ear Acupuncture in the Treatment of Migraine Attacks: A Randomized Trial on the Efficacy of Appropriate Versus Inappropriate Acupoints." *Neurological Sciences* 32 (1): 173–75. https://doi.org/10.1007/s10072-011-0525-4.

Allison, T., and D. V. Cicchetti. 1975. "Sleep in Mammals: Ecological and Constitutional Correlates." *Arch. Hydrobiol* 75: 442. https://doi.org/10.1126/science.982039.

American Council on Education. 2008. "College-Bound Students' Interests in Study Abroad and Other International Learning Activities." http://www.openintro.org/redirect.php?go=textbook-Interests_in_Study_Abroad_2008.

Anturane Reinfarction Trial Research Group. 1980. "Sulfinpyrazone in the Prevention of Sudden Death After Myocardial Infarction." *New England Journal of Medicine* 302 (5): 250–56. https://doi.org/10.1056/NEJM198001313020502.

Asbury, D. A., and S. C. Adolph. 2007. "Behavioral Plasticity in an Ecological Generalist: Microhabitat Use by Western Fence Lizards." *Evolutionary Ecology Research* 9: 801–15.

Audera, C., R. V. Patulny, Sander B. H, and R. M. Douglas. 2001. "Mega-Dose Vitamin c in Treatment of the Common Cold: A Randomised Controlled Trial." *Medical Journal of Australia* 175 (7): 359–62. http://www.openintro.org/redirect.php?go=textbook-vitamin_C_cold_treatment_2001.

Backstrom, L. 2011. "Anatomy of Facebook." *Facebook Data Team's Notes*. http://www.openintro.org/redirect.php?go=textbook-anatomy-of-facebook.

Benson, J. B. 1993. "Season of Birth and Onset of Locomotion: Theoretical and Methodological Implications." *Infant Behavior and Development* 16 (1): 69–81. https://doi.org/10.1016/0163-6383(93)80029-8.

Bertrand, M., and S. Mullainathan. 2003. "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." https://doi.org/10.3386/w9873.

Böttiger, B. W., C. Bode, S. Kern, A. Gries, R. Gust, R. Glätzer, H. Bauer, J. Motsch, and E. Martin. 2001. "Efficacy and Safety of Thrombolytic Therapy After Initially Unsuccessful Cardiopulmonary Resuscitation: A Prospective Clinical Trial." *The Lancet* 357 (9268): 1583–85.

Bucciol, Alessandro, and Marco Piovesan. 2011. "Luck or Cheating? A Field Experiment on Honesty with Children." *Journal of Economic Psychology* 32 (1): 73–78. https://doi.org/10.1016/j.joep.2010.12.001.

CDC. 2008. "Perceived Insufficient Rest or Sleep Among Adults – United States, 2008." http://www.openintro.org/redirect.php?go=textbook-Perceived_Insufficient_Rest_or_Sleep_Among_Adults.

———. 2018. "2018 Assisted Reproductive Technology Fertility Clinic Success Rates Report." https://www.cdc.gov/art/pdf/2018-report/ART-2018-Clinic-Report-Full.pdf.

Chance, B., and A. Rossman. 2018. *Investigating Statistical Concepts, Applications, and Methods*.

http://www.rossmanchance.com/iscam3/.

Chimowitz, M. I., M. J. Lynn, C. P. Derdeyn, T. N. Turan, D. Fiorella, B. F. Lane, L. S. Janis, et al. 2011. "Stenting Versus Aggressive Medical Therapy for Intracranial Arterial Stenosis." *New England Journal of Medicine* 365 (11): 993–1003. http://www.nejm.org/doi/full/10.1056/NEJMoa1105335.

Conner, T. S., K. L. Brookie, A. C. Carr, L. A. Mainvil, and M. CM. Vissers. 2017. "Let Them Eat Fruit! The Effect of Fruit and Vegetable Consumption on Psychological Well-Being in Young Adults: A Randomized Controlled Trial." *PloS One* 12 (2). https://doi.org/10.1371/journal.pone.0171206.

Datoo, M. S., M. H. Natama, A. Somé, O. Traoré, T. Rouamba, D. Bellamy, P. Yameogo, et al. 2021. "High Efficacy of a Low Dose Candidate Malaria Vaccine, R21 in 1 Adjuvant Matrix-m^{TM}, with Seasonal Administration to Children in Burkina Faso." *The Lancet*. https://doi.org/10.1016/S0140-6736(21)00943-0.

Demos. 2011. "The State of Young America: The Poll." http://www.openintro.org/redirect.php?go=textbook-young_americans_2011_extra.

Ellis, G. J., and L. H. Stone. 1979. "Marijuana Use in College:" an Evaluation of a Modeling Explanation"." *Youth and Society* 10 (4): 323.

FiveThirtyEight. 2015. http://www.openintro.org/redirect.php?go=textbook-fivethirtyeight-scary-movies.

Frederick, S., N. Novemsky, J. Wang, R. Dhar, and S. Nowlis. 2009. "Opportunity Cost Neglect." *Journal of Consumer Research* 36 (4): 553–61.

Gallup. 2012. http://www.openintro.org/redirect.php?go=textbook-employed_americans_in_better_health_2012.

Gallup. 2021a. "Half of College Students Say COVID-19 May Impact Completion." https://www.openintro.org/go?id=textbook-gallup-2021-covid-college-impact.

———. 2021b. "U.s. Support for Vaccination Proof Varies by Activity, Data Collected in April 2021." https://www.openintro.org/go?id=textbook-gallup-2021-vaccine-proof.

Garbutt, J. M., C. Banister, E. Spitznagel, and J. F. Piccirillo. 2012. "Amoxicillin for Acute Rhinosinusitis: A Randomized Controlled Trial." *JAMA: The Journal of the American Medical Association* 307 (7): 685–92. https://doi.org/10.1001/jama.2012.138.

Gneezy, Uri, and Aldo Rustichini. 2000. "A Fine Is a Price." *The Journal of Legal Studies* 29 (1): 1–17. https://doi.org/10.1086/468061.

Gorman, K. B., T. D. Williams, and W. R. Fraser. 2014b. "Ecological sexual dimorphism and environmental variability within a community of Antarctic penguins (genus Pygoscelis)." *PloS One* 9 (3): e90081. https://doi.org/doi.org/10.1371/journal.pone.0090081.

———. 2014a. "Ecological Sexual Dimorphism and Environmental Variability Within a Community of Antarctic Penguins (Genus Pygoscelis)." *PLoS ONE* 9(3) (e90081): –13. https://doi.org/10.1371/journal.pone.0090081.

Graham, D. J., R. Ouellet-Hellstrom, T. E. MaCurdy, F. Ali, C. Sholley, C. Worrall, and J. A. Kelman. 2010. "Risk of Acute Myocardial Infarction, Stroke, Heart Failure, and Death in Elderly Medicare Patients Treated with Rosiglitazone or Pioglitazone." *JAMA* 304 (4): 411. https://doi.org/10.1001/jama.2010.920.

Hand, D. J. 1994. *A handbook of small data sets*. Chapman & Hall/CRC.

Hayden, R. W. 2019. "Questionable Claims for Simple Versions of the Bootstrap." *Journal of Statistics Education* 27 (3): 208–15. https://doi.org/10.1080/10691898.2019.1669507.

Heinz, G., L. J. Peterson, R. W. Johnson, and C. J. Kerk. 2003. "Exploring Relationships in Body Dimensions." *Journal of Statistics Education* 11 (2). http://www.openintro.org/redirect.php?go=textbook-body_dim_2003.

Hepler, J., and D. Albarracín. 2013. "Attitudes Without Objects: Evidence for a Dispositional Attitude, Its Measurement, and Its Consequences." *Journal of Personality and Social Psychology* 104 (6): 1060. https://doi.org/10.1037/a0032282.

Hesterberg, T. C. 2015. "What Teachers Should Know about the Bootstrap: Resampling in the Undergraduate Statistics Curriculum." *The American Statistician* 69 (4): 371–86. https://doi.org/10.1080/00031305.2015.1089789.

ICPSR. 2014. "United States Department of Health and Human Services. Centers for Disease Control and Prevention. National Center for Health Statistics. Natality Detail File, 2014 United States. Inter-University Consortium for Political and Social Research, 2016-10-07." https://doi.org/10.3886/ICPSR36461.v1.

Kaiser Family Foundation. 2019. https://www.openintro.org/go?id=textbook-kff-2019-expand-aca-coverage.

King, C., J. Suamani, N. Sanuku, Y.-C. Cheng, S. Satofan, B. Mancuso, C. W. Goss, et al. 2018. "A Trial of a Triple-Drug Treatment for Lymphatic Filariasis." *New England Journal of Medicine* 379: 1801–10. https://doi.org/10.1056/NEJMoa1706854.

Latter, O. H. 1902. "The Egg of Cuculus Canorus. An Enquiry into the Dimensions of the Cuckoo's Egg and the Relation of the Variations to the Size of the Eggs of the Foster-Parent, with Notes on Coloration." *Biometrika* 1: 164–76. https://doi.org/10.2307/2331486.

Legaki, Nikoletta-Zampeta, Nannan Xi, Juho Hamari, Kostas Karpouzis, and Vassilios Assimakopoulos. 2020. "The Effect of Challenge-Based Gamification on Learning: An Experiment in the Context of Statistics Education." *International Journal of Human-Computer Studies* 144. https://doi.org/10.1016/j.ijhcs.2020.102496.

Lo Sasso, A. T., M. R. Richards, C. F. Chou, and S. E. Gerber. 2011. "The $16,819 Pay Gap for Newly Trained Physicians: The Unexplained Trend of Men Earning More Than Women." *Health Affairs* 30 (2). https://doi.org/10.1377/hlthaff.2010.0597.

Lockman, S., R. L. Shapiro, L. M. Smeaton, C. Wester, I. Thior, L. Stevens, F. Chand, et al. 2007. "Response to Antiretroviral Therapy After a Single, Peripartum Dose of Nevirapine." *New England Journal of Medicine* 356 (2): 135–47. https://doi.org/10.1056/NEJMoa062876.

Lucas, M., F. Mirzaei, A. Pan, O. I. Okereke, W. C. Willett, E. J. O'Reilly, K. Koenen, and A. Ascherio. 2011. "Coffee, Caffeine, and Risk of Depression Among Women." *Archives of Internal Medicine* 171 (17): 1571. https://doi.org/10.1001/archinternmed.2011.393.

Malkevitch, J., and L. M. Lesser. 2008. *For All Practical Purposes: Mathematical Literacy in Today's World*. WH Freeman & Co.

Manson, J. E., N. R. Cook, I-M. Lee, W. Christen, S. S. Bassuk, S. Mora, H. Gibson, et al. 2019. "Marine n-3 Fatty Acids and Prevention of Cardiovascular Disease and Cancer." *New England Journal of Medicine* 380 (1): 23–32. https://doi.org/10.1056/NEJMoa1811403.

McNeil, D. R. 1977. "Interactive Data Analysis: A Practical Primer."

Ménard, C., A. A. Hagège, O. Agbulut, M. Barro, M. C. Morichetti, C. Brasselet, A. Bel, et al. 2005. "Transplantation of Cardiac-Committed Mouse Embryonic Stem Cells to Infarcted Sheep Myocardium: A Preclinical Study." *The Lancet* 366 (9490): 1005–12.

Mortada, WI, MA Sobh, MM El-Defrawy, and SE Farahat. 2000. "Study of Lead Exposure from Automobile Exhaust as a Risk for Nephrotoxicity Among Traffic Policemen." *American Journal of Nephrology* 21 (4): 274–79. https://doi.org/10.1159/000046261.

National Sleep Foundation. 2012. http://www.openintro.org/redirect.php?go=textbook-trans_workers_sleep_2012.

Nieman, D. C., E. J. Cayea, M. D. Austin, D. A. Henson, S. R. McAnulty, and F. Jin. 2009. "Chia Seed Does Not Promote Weight Loss or Alter Disease Risk Factors in Overweight Adults." *Nutrition Research* 29 (6): 414–18. https://doi.org/10.1016/j.nutres.2009.05.011.

Nightingale, Florence. 1859. *Notes on Nursing: What It Is, and What It Is Not*. Harrison & Sons.

NOAA. 2023. "Climate Data Online." http://www.openintro.org/redirect.php?go=textbook-noaa_1948_2018.

NORC. 2010. http://www.openintro.org/redirect.php?go=textbook-gss_2010.

———. 2016. "Data Explorer - Government Resp to Promote Gender Equality." https://www.openintro.org/go?id=textbook-gssdataexplorer-marijuana-gender-equality.

———. 2022. "Data Explorer - Should Marijuana Be Made Legal." https://www.openintro.org/go?id=textbook-gssdataexplorer-marijuana-legal.

Nuñez, J. R., C. R. Anderton, and R. S. Renslow. 2018. "Optimizing Colormaps with Consideration for Color Vision Deficiency to Enable Accurate Interpretation of Scientific Data." *PloS One* 13 (7). https://doi.org/10.1371/journal.pone.0199239.

O'Connor, A. 2011. "Coffee Drinking Linked to Less Depression in Women." *New York Times*. http://www.openintro.org/redirect.php?go=textbook-coffee_depression_2011.

Oldham-Cooper, R. E., C. A. Hardman, C. E. Nicoll, P. J. Rogers, and J. M. Brunstrom. 2011. "Playing a Computer Game During Lunch Affects Fullness, Memory for Lunch, and Later Snack Intake." *The American Journal of Clinical Nutrition* 93 (2): 308. https://doi.org/10.3945/ajcn.110.004580.

Orben, A., and A. K. Baukney-Przybylski. 2018. "Screens, Teens and Psychological Well-Being: Evidence from Three Time-Use Diary Studies." *Psychological Science*. https://doi.org/10.1177/0956797619830329.

Othman, S., N. Hassan, and A. Mohamed. 2023. "Effectiveness of Mindfulness-Based Interventions on Burnout and Self-Compassion Among Critical Care Nurses Caring for Patients with COVID-19: A Quasi-Experimental Study." *BMC Nursing* 22: 35.

Parker-Pope, T. 2011. "The School Bully Is Sleepy." *New York Times*. http://www.openintro.org/redirect.php?go=textbook-school_bully_sleepy_2011.

Pew Research Center. 2011. "Is College Worth It?, Data Collected March 15-29, 2011." http://www.openintro.org/redirect.php?go=textbook-college_worth_it_2011.

———. 2013. "The Diagnosis Difference." https://www.openintro.org/go?id=textbook-pew-2013-diagnosis-difference.

———. 2018. "A Majority of Teens Have Experienced Some Form of Cyberbullying, Data Collected March 7 - April 10, 2018." https://www.openintro.org/go?id=textbook-pew-2018-cyberbullying.

———. 2021a. "How Americans' Attitudes about Climate Change Differ by Generation, Party and Other Factors, Data Collected April 20 to 29, 2021." https://www.openintro.org/go?id=textbook-pew-2021-climate-change.

———. 2021b. "The State of Online Harassment, Data Collected September 8 - 13, 2020." https://www.openintro.org/go?id=textbook-pew-2021-online-harassment.

———. 2022. "News Consumption Across Digital Platforms." https://www.pewresearch.org/journalism/fact-sheet/news-platform-fact-sheet/pj_2022-09-20_news-platforms_0-03b/.

Pfizer. 2021. "Pfizer-Biontech Announce Positive Topline Results of Pivotal Covid-19 Vaccine Study in Adolescents." https://www.openintro.org/go?id=textbook-pfizer-2021-adolescent-covid-vaccine.

Pierson, Emma, Camelia Simoiu, Jan Overgoor, Sam Corbett-Davies, Daniel Jenson, Amy Shoemaker, Vignesh Ramachandran, et al. 2020. "A Large-Scale Analysis of Racial Disparities in Police Stops Across the United States." *Nature Human Behaviour* 4 (7): 736–45. https://doi.org/10.1038/s41562-020-0858-1.

Piff, P. K., D. M. Stancato, S. Côté, R. Mendoza-Denton, and D. Keltner. 2012. "Higher Social Class Predicts Increased Unethical Behavior." *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1118373109.

Poll, Marist. 2011. "Road Rules: Re-Testing Drivers at Age 65?, March 4, 2011." http://www.openintro.org/redirect.php?go=textbook-drivers_at_65_2011.

Polling, Public Policy. 2015. "Americans on College Degrees, Classic Literature, the Seasons, and More, Data Collected Feb 20-22, 2015." http://www.openintro.org/redirect.php?go=textbook-PPP_30215.

Rabin, R. C. 2010. "Risks: Smokers Found More Prone to Dementia." *New York Times*. http://www.openintro.org/redirect.php?go=textbook-nytimes_smoking_dementia.

Ramsey, F., and D. Schafer. 2012. *The Statistical Sleuth*. 3rd ed. Cengage Learning.

Rao, M., L. Bar, Y. Yu, M. Srinivasan, A. Mukherjea, J. Li, S. Chung, S. Venkatraman, S. Dan, and L. Palaniappan. 2021. "Disaggregating Asian American Cigarette and Alternative Tobacco Product Use: Results from the National Health Interview Survey (NHIS) 2006–2018." *Journal of Racial and Ethnic Health Disparities*. https://doi.org/10.1007/s40615-021-01024-5.

Ritz, B., F. Yu, G. Chapa, and S. Fruin. 2000. "Effect of Air Pollution on Preterm Birth Among Children Born in Southern California Between 1989 and 1993." *Epidemiology* 11 (5): 502–11. http://www.jstor.org/stable/3703990.

Rosen, B., and T. H. Jerdee. 1974. "Influence of Sex Role Stereotypes on Personnel Decisions." *Journal of Applied Psychology* 59 (1): 9. https://doi.org/doi.org/10.1037/h0035834.

Rubio-Fernandez, P., F. Mollica, and J. Jara-Ettinger. 2021. "Speakers and Listeners Exploit Word Order for Communicative Efficiency: A Cross-Linguistic Investigation." *Journal of Experimental Psychology: General* 150 (3): 583–94. https://doi.org/10.1037/xge0000963.

Scanlon, T. J., R. N. Luben, F. L. Scanlon, and N. Singleton. 1993. "Is Friday the 13th Bad for Your Health?" *British Medical Journal* 307: 1584–86. https://doi.org/10.1136/bmj.307.6919.1584.

Smith, G. E., P. A. Chouinard, and S. Byosiere. 2021. "If i Fits i Sits: A Citizen Science Investigation into Illusory Contour Susceptibility in Domestic Cats (*Felis Silvestris Catus*)." *Applied Animal Behaviour Science* 240. https://doi.org/10.1016/j.applanim.2021.105338.

Spear, M. E. 1952. *Charting Statistics*. McGraw Hill Book Compay, Inc.

Survey USA. 2012. "News Poll 19333, Data Collected on June 27, 2012." http://www.openintro.org/redirect.php?go=textbook-SurveyUSA_19333.

———. 2019. http://www.openintro.org/redirect.php?go=textbook-SurveyUSA_24568.

———. 2021. https://www.openintro.org/go?id=textbook-surveyusa-2021-seattle-defund-police.

Turnbull, B., B. Brown, and M. Hu. 1974. "Survivorship of Heart Transplant Data." *Journal of the*

American Statistical Association 69: 74–80. https://doi.org/10.1080/01621459.1974.10480130.

UNC Carolina Population Center. 2006. "China Health and Nutrition Survey, 2006." http://www.openintro.org/redirect.php?go=textbook-china_health_nut_survey_2006.

US Census Bureau. 2010. "2010 Census Urban and Rural Classification and Urban Area Criteria Housing Characteristics." http://www.openintro.org/redirect.php?go=textbook-housing_char_2010.

US DOE EPA. 2021. "Fuel Economy, 2021 Data File." https://www.openintro.org/go?id=textbook-fueleconomy-2021.

Vaughn, A. 2011. "Poll Finds Young Adults Optimistic, but Not about Money." *Los Angeles Times*. http://www.openintro.org/redirect.php?go=textbook-young_americans_2011.

Washington Post. 2020. "Washington Post-Schar School National Poll, Data Collected June 2-7, 2020." http://www.openintro.org/redirect.php?go=exercise-wapo-2020-blm-poll.

Wickham, H. 2014. "Tidy Data." *Journal of Statistical Software* 59 (10). https://doi.org/10.18637/jss.v059.i10.

———. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Witmer, J. 2021. "Simpson's Paradox, Visual Displays, and Causal Diagrams." *The American Mathematical Monthly* 128 (7): 598–610. https://doi.org/10.1080/00029890.2021.1932237.

You, W., R. Henneberg, A. Saniotis, Y. Ge, and M. Henneberg. 2022. "Total Meat Intake Is Associated with Life Expectancy: A Cross-Sectional Data Analysis of 175 Contemporary Populations." *International Journal of General Medicine* 15: 1833–51.

YouGov. 2021. "Do You Think the Coronavirus Will End up Bringing the World Closer Together, or Leave Us Further Apart?" https://www.openintro.org/go?id=textbook-yougov-2021-covid.

Index

95% confidence interval, 229
95% confident, 229

accuracy, 463
adjusted R-squared, 143
AIC, 159
Akaike information criterion, 159
alternative hypothesis, 181, 188
Analysis of Variance, 378
analysis of variance (ANOVA), 378
anecdotal evidence, 29
ANOVA, 378
    degrees of freedom, 382
association, 16, 17
average, 74

backward elimination, 144
bias, 30
    non-response, 31
bimodal, 78
blind, 36
blinding, 36
blocking, 34
boostrap percentile confidence interval, 205
bootstrap
    CI paired difference, 365
    percentile interval, 284
    SE interval, 284
bootstrap CI for the slope, 414
bootstrap CI paired difference, 365
bootstrap sample, 205
bootstrapping, 198, 199
box plot, 81

case, 12
census, 28
Central Limit Theorem, 214, 325, 327
Chi-squared distribution, 307
chi-squared statistic, 306
cluster sample, 32
coefficient of determination, 120
cohort, 18
collinear, 142
column proportions, 60
column totals, 56
conditional proportions, 60
confidence interval, 197, 205, 229, 231, 264
confidence level, 264
confirmation bias, 241
confounding variable, 34
contingency table, 56
control, 34
control group, 35
convenience sample, 31
correlation, 111
cross-validation, 442, 463

data density, 77
data fishing, 381
data frame, 13
data snooping, 381
decision errors, 239, 240
degrees of freedom, 143, 328, 382
    ANOVA, 382
    chi-squared test, 308
    single mean, 328
density plot, 77
dependent variable, 17
deviation, 78
difference in means, 346
    point estimate, 342, 351
    standard error (SE), 351
    T score, 350
difference in proportions
    standard error (SE), 286
    Z score, 289
difference of proportions
    point estimate, 279
discernibility level, 190, 240
distribution, 74
dodged bar plot, 57
dot plot, 74
double-blind, 36

expected counts, 304
experiment, 18, 34
    placebo, 36
explanatory variable, 17, 438
extrapolation, 119

F-test, 386
faceted plot, 64
filled bar plot, 57
first quartile, 82
forward selection, 144
full model, 143

generalized linear model, 155

high leverage, 124
histogram, 76
hypothesis test, 178, 181, 188
hypothesis testing, 241, 245
  discernibility level, 190
  p-value, 189
  statistically discernible., 190

independence, 17, 303
independent, 17
independent variable, 17, 179
indicator variable, 121, 472
influential point, 124
intensity map, 86
interquartile range, 82
IQR, 82

least squares line, 116
left skewed, 77
leverage point, 124
logistic regression, 155, 461
logit transformation, 157
lurking variable, 34

margin of error, 225, 226
mean, 74
mean square between groups, 382
mean square error, 382
median, 81
mosaic plot, 58
multicollinearity, 142, 441, 474
multimodal, 78
multiple comparisons, 387
multiple predictors, 438, 462
multiple regression, 140
multistage sample, 32

non-response rate, 31
nonlinear, 73
normal curve, 215
normal distribution, 215
normal model, 215
normal probability table, 218
null distribution, 213, 242
null hypothesis, 181, 188
numerical data, 319

observation, 12
observational study, 18, 37
one-sided hypothesis test, 241
outcome variable, 106
outlier, 82, 124

p-value, 189
paired data, 360, 365
paired difference
  T score, 368
parameter, 199, 216
parametric bootstrap, 259
parsimonious, 143
percentile, 82, 218
percentile interval, 284
permutation test, 181
placebo, 18, 36
placebo effect, 36
plot
  box, 81
  density, 77
  dodged bar, 57
  dot, 74
  faceted, 64
  filled bar, 57
  histogram, 76
  intensity map, 86
  mosaic, 58
  ridge, 64
  scatterplot, 72
  side-by-side box, 63
  stacked bar, 57
  standardized bar, 57
point estimate, 75, 181, 199, 324
  difference in means, 342, 351
  difference of proportions, 279
  single mean, 319, 325
  single proportion, 262
pooled proportion, 288
population, 28, 30
population parameter, 29
power, 245
prediction error, 442
predictor variable, 106, 438
probability of an event, 157
prominent, 78
prospective study, 37
pseudoreplication, 34

R-squared, 119
randomization test, 179, 191
randomization test for the slope, 412
randomized experiment, 18, 34
reference level, 138
regression sum of squares, 120
replication, 34
replication crisis, 34
representative sample, 31
residuals, 109
response variable, 17
retrospective study, 37
ridge plot, 64
right skewed, 77
robust statistic, 84
row proportions, 60
row totals, 56

sample, 28, 30
sample bias, 30
sample statistic, 29, 83
sampling

bias, 30
cluster, 32
convenience, 31
multistage, 32
random, 30
representative, 31
simple random, 30, 31
strata, 32
stratified, 32
sampling distribution, 212
sampling with replacement, 205
scatterplot, 16, 72
SE interval, 284
side-by-side box plot, 63
significance level, 190, 240
simple random sample, 30, 31
Simpson's paradox, 51
simulation, 182
single mean
point estimate, 325
standard error (SE), 325
T score, 333
single proportion
point estimate, 262
standard error (SE), 262
Z score, 266
stacked bar plot, 57
standard deviation, 79, 323
standard error, 225, 323
standard error (SE), 323
difference in means, 351
difference in proportions, 286
single mean, 325
single proportion, 262
standard normal distribution, 215
standardized bar plot, 57
statistic, 183, 199
statistical inference, 184
statistically discernible, 190
statistically significant, 190
stepwise selection, 143
stratified sample, 32
strong skew, 82
study
cohort, 18
experiment, 18, 34
observational, 18, 37
prospective, 37
randomized experiment, 18, 34
retrospective, 37
success, 186
success-failure condition, 262
sum of squared errors, 120, 382
sum of squares between groups, 382
sum of squares total, 382
summary statistic, 11, 15, 83
symmetric, 77

T score, 333
difference in means, 350
paired difference, 368
single mean, 333
slope, 419
t-CI
difference in means, 351
paired difference, 369
single mean, 332
t-distribution, 327, 330
t-distribution, 325
t-test, 333, 350
tail, 77
technical conditions linear regression, 422
technical conditions logistic regression, 460
test statistic, 183, 189
third quartile, 82
tidy data, 13
total sum of squares, 120
transformation, 85, 157
treatment group, 35
two-sided hypothesis test, 241
Type I error, 239
Type II error, 239

unimodal, 78
unit of observation, 12

variability, 78, 82
variable, 12
categorical, 14
continuous, 14
dependent, 17
discrete, 14
explanatory, 17, 438
independent, 17, 179
indicator, 121, 472
levels, 14
multiple predictors, 438, 462
nominal, 14
numerical, 14
ordinal, 14
outcome, 106
predictor, 106, 438
response, 17
variance, 79

weighted mean, 76
whisker, 82

Z score, 217, 266
difference in proportions, 289
single proportion, 266
Z-CI
difference in proportions, 286
single proportion, 264